COMPOSITE LIKELIHOOD AND REGRESSION BASED METHODS FOR INFERRING POPULATION GENETIC PARAMETERS FROM DNA SEQUENCES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Lan Zhu

August 2006

COMPOSITE LIKELIHOOD AND REGRESSION BASED METHODS FOR INFERRING POPULATION GENETIC PARAMETERS FROM DNA SEQUENCES

Lan Zhu, Ph.D.

Cornell University, 2006

This doctoral dissertation is composed of three projects and four chapters. Chapter 1 presents the background theories and models in the field of population genetics that are related to these projects.

In Chapter 2, I develop a composite likelihood ratio test (CLRT) for detecting genes and genomic regions that are subject to recurrent natural selection while relaxing the assumption of free recombination. We find that the test has excellent power to detect weak negative selection and moderate power to detect positive selection. Moreover, the test is quite robust to the bias in the estimate of local recombination rate, but not to certain demographic scenarios such as population growth or a recent bottleneck.

In Chapter 3, I present a novel method, Poisson pairwise difference method (PPDM), which efficiently co-estimates the selection coefficient $\gamma = 4N_e s$ and mutation rate $\theta = 4N_e \mu$ from arbitrarily correlated SFS data. We demonstrate that the PPDM log-likelihood ratio test has good power to detect positive selection and moderate power to detect weak negative selection.

Current state-of-the-art approaches for quantifying meiotic recombination rates (R) and/or identifying hotspots are mostly based on the likelihood of observed

haplotypes or linkage disequilibrium (LD) patterns. In Chapter 4, I describe a flexible, efficient, and population structure robust approach via multiple linear regression and non-parametric bootstrap based on the frequency spectra of unphased single nucleotide polymorphism sites (SNPs) and provide confidence intervals of R between adjacent pairs of SNPs. No LD information is required. We evaluate this new approach via Monte Carlo simulation as well as application to the well-characterized hotspots near the human $TAP2$ gene and a 206-kb region on $ch1q42.3$ near $MS32$.

BIOGRAPHICAL SKETCH

Lan Zhu earned her Bachelor of Medicine degree from Beijing Medical University (now Peking University Health Science Center, China) in 1996. After graduation she began work in the Department of Environment Health, School of Public Health, in Beijing Medical University. In 2001 she started her graduate study in the field of Biological Statistics and Computational Biology at Cornell University. She earned her Master of Science in Biometry in 2004.

To my family

ACKNOWLEDGEMENTS

First, I would like to thank my advisor, Dr. Carlos Bustamante, for his unfailing encouragement and direction during my graduate study. As a mentor and as a scholar, Carlos's guidance and enthusiasm have been invaluable to me. This work could not have been finished without his effort.

I would like to thank the other members of my committee: Martin Wells and Charles Aquadro, for their time and valuable comments.

I also want to thank Jaroslaw Pillardy in the Cornell Theory Center for helping me to get my programs to run correctly on the CTC cluster, and James VanEe and Jason Allen in BRC Computing Facility for helping me to install softwares and solve computer problems. I really appreciate their constant help.

I am grateful to Russ Lloyd for his kindness and enthusiasm. We had a great time when I was TA for his BTRY601/602 classes. The ice fishing morning with him is such a joyful and beautiful scene that is always in my memory of Ithaca.

I also would like to thank Dr. Edward Gunn, who accepted me as a Teaching Assistant in the department of Asian Studies so that I could have the financial support for my first year graduate study.

Many thanks to all members in the Bustamante lab and students, faculty and staff in the department of Biological Statistics and Computational Biology for their kind help and warm friendship.

Lastly, I would like to thank my husband, Feng Feng, for his understanding, computer programming assistance, detailed project discussion and wonderful comments. I would also like to thank my parents and parents-in-law for their encouragement and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CLRT, composite likelihood ratio test

FISHER, forward infinite-sites simulation with selection and recombination

GOF, goodness-of-fit test

LD, linkage disequilibrium

MCLE, maximum composite likelihood estimator

MLE, maximum likelihood estimator

MLR, multiple linear regression

PPDM, Poisson pairwise difference method

PPDSFS, Poisson pairwise difference site frequency spectrum

PRF, Poisson random field

SFS, site frequency spectrum

SNPs, single nucleotide polymorphism site

Chapter 1

Introduction

A central goal of population genetics is to understand the evolutionary forces that produce and maintain genetic variation within populations and between species. These forces include mutation, recombination, genetic drift, natural selection, and population structure. A major theme of current research in population genetics has been the development of statistical methods for comparing different evolutionary models for a given data set. For example, one may wish to ask whether a region of DNA sequenced across a random sample of individuals from the same population shows signatures of recent natural selection or variation in recombination rate. Coalescent theory, introduced by Kingman [44, 45, 46] and also discovered independently by Hudson [27] and by Tajima [78], is a fundamental tool in the study of genetic sequence variation that probabilistically describes the mathematical process of joining up sampled sequences into common ancestor. The purpose of this chapter is to introduce one of the most commonly used models in population genetics, the Wright-Fisher model, which in the limit of large population size can be approximated by Kingman's [44, 45, 46] coalescent, reviewed by Donnelly and Tavaré (1995)[10] and Nordborg(2001) [67]. We will then introduce the Poisson Random Field (PRF) setting, which forms the basis of our project for estimating mutation and selection parameters under various assumptions regarding the rate of recombination.

1.1 *The Wright-Fisher Model*

Consider a diploid population of constant size N (that is, $2N$ chromosomes) that reproduces in discrete non-overlapping generations. The mating scheme is random, that is, each individual is likely to mate with every another individual. Reproduction is a random process in the sense that individuals may or may not contribute any offspring to the next generation. This stochastic process by which gametes are sampled will be modeled using the Binomial distribution (genetic drift). The population of the present generation is obtained by random sampling with replacement from the previous generation. Let us focus on one neutrally evolving locus with two alleles A and a segregating in the population, and assume that there is no mutation, no difference in fitness between these two alleles, and no population structure. Let Y_t be the random variable describing the number of copies of allele A at generation t in the population. At time $t = 0$, there are $x(0 \leq x \leq 2N)$ copies of allele A and $2N - x$ copies of allele a in the population. Thus $p_t = y_t/2N$ is the frequency of A at generation t in the population. Because the population is finite in size and N is constant over time, under the above assumptions, the random variable $Y_{t+1} = y_{t+1}|Y_t = y_t$ follows a binomial distribution with parameter $(2N, p_t)$ with density formula,

$$P(Y_{t+1} = y_{t+1}|Y_t = y_t) = \begin{pmatrix} 2N \\ y_{t+1} \end{pmatrix} p_t^{y_{t+1}}(1 - p_t)^{(2N - y_{t+1})} \tag{1.1}$$

This probability is called the transition probability of the stochastic process. It gives the probability that a gene with y_t copies in the present generation is found in y_{t+1} copies in next generation. The mean and variance of number of copies of allele A in generation $t + 1$ given it is y_t at generation t comes directly from the

binomial distribution,

$$E(Y_{t+1}|Y_t = y_t) = 2N \times \frac{y_t}{2N} = y_t \tag{1.2}$$

$$Var(Y_{t+1}|Y_t = y_t) = 2N \times \frac{y_t}{2N} \times (1 - \frac{y_t}{2N}) = 2N \times p_t(1 - p_t) \tag{1.3}$$

The model is known as the Wright-Fisher model introduced by Fisher (1930) [18] and Wright (1931) [89]. One property that can be immediately seen in the Wright-Fisher model is that if all assumptions are satisfied, the expected change in allele frequency from generation to generation is zero.

1.2 *The Standard Coalescent Model*

A coalescent is a stochastic process which models the ancestry of a random sample of n individuals from a population that has evolved with constant size N over many generations. This process was described by Kingman [44, 45, 46], and also Hudson [27] and Tajima [78]. It was proved to be limiting ancestral process for a wide variety of neutral demographic models that includes the Wright-Fisher model. Here we will briefly introduce some concepts of the coalescent process in the Wright-Fisher model that are related to our projects, such as the *time to the most recent common ancestor (MRCA)*, *the number of segregating sites*, and *the site-frequency spectrum(SFS)*.

1.2.1 *Time to the Most Recent Common Ancestor (MRCA)*

Under the Wright-Fisher model, it is possible to describe the probability distribution of coalescence times. For n sampled individuals at a locus from a population of size N, if we consider each individual at that locus as a lineage, we can use the

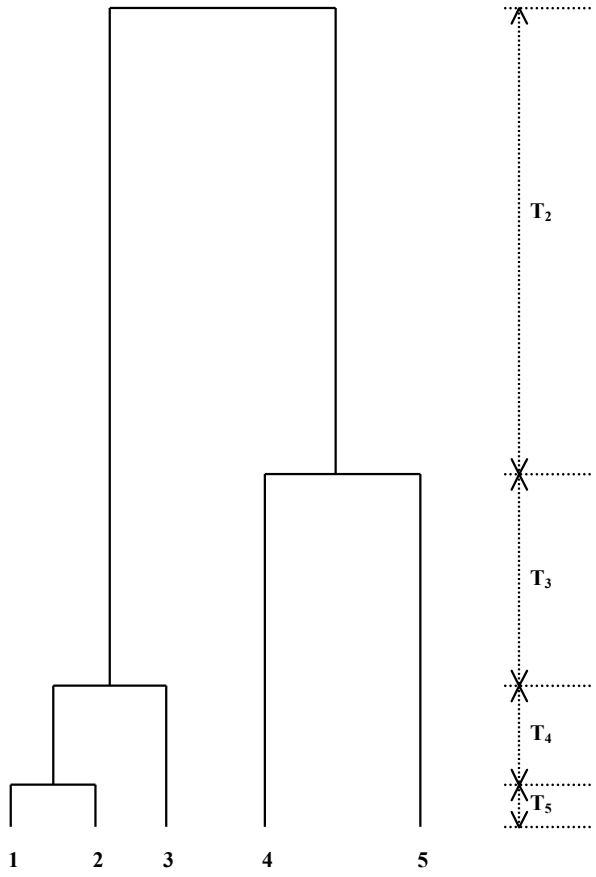


Figure 1.1: *Time to the most recent common ancestor*

coalescent to trace the ancestral lineages back through time. Suppose at each coalescent event, only two of the lineages fuse into one common ancestral lineage. So for a sample of n lineages at the present time, the first coalescent event decreases the number of lineages from n to $n-1$, then second one decreases it from $n-1$ to $n-2$, etc, until the last coalescent event results the most recent common ancestor (MRCA). Define a random variable T_i to be the time that it takes for a coalescent event such that the number of lineages decreases from i to $i-1$. As shown in the figure 1.1. Kingman [44, 45] showed that in the limit of large population (as N goes to infinity), when time is measured appropriately, the coalescent time (T_n, T_{n-1}, ..., T_2) are independent and exponentially distributed with densities and expected values,

$$f_{T_i}(t_i) = \begin{pmatrix} i \\ 2 \end{pmatrix} e^{-\begin{pmatrix} i \\ 2 \end{pmatrix} t} \tag{1.4}$$

$$E[T_i] = \frac{2}{i(i-1)} \tag{1.5}$$

Since the $T_i's$ are independent, we can get the mean time to the most recent common ancestor directly from the sum of the expected value of the individual $T_i's$,

$$\begin{aligned} E[T_{MRCA}] &= \sum_{i=2}^{n} E[T_i] \\ &= 2\sum_{i=2}^{n}(\frac{1}{i-1} - \frac{1}{i}) \\ &= 2(1 - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} + \frac{1}{3} - \ldots - \frac{1}{n-1} + \frac{1}{n-1} - \frac{1}{n}) \\ &= 2(1 - \frac{1}{n}) \end{aligned} \tag{1.6}$$

Similarly, the total length of all lineages for a sample of size n is,

$$E[T_{total}] = \sum_{i=2}^{n} iE[T_i] = \sum_{i=2}^{n} i\frac{2}{i(i-1)} = 2\sum_{i=2}^{n-1}\frac{1}{i} \tag{1.7}$$

1.2.2 *Number of Segregating Sites*

Assume that mutations enter the population as a poisson process with rate $\theta/2 = 2N_e\mu$, where N_e is the effective population size, μ is the mutation rate per locus per generation. Then for a branch(lineage) with length T_i, the expected number of mutants at one locus is

$$E[S_i] = \theta/2 \times E[T_i] \tag{1.8}$$

Under the infinite sites model, that is, when each mutation occurs at a previously invariant DNA site, for n sampled sequences with k loci, the expected number of total mutations, which is defined as the number of segregating sites of the interested region would be,

$$E[S] = \theta/2 \times E[T_{total}] \times k = \theta/2 \times 2\sum_{i=2}^{n-1}\frac{1}{i} \times k = \theta k\sum_{i=2}^{n-1}\frac{1}{i} \tag{1.9}$$

1.3 *The Poisson Random Field*

Define γ to be the selection parameter and let $\gamma = 2N_e s$, where s is the fitness effect of new mutations such that wild-type fitness is 1, heterozygote fitness is $1+s$, and homozygote fitness for the new mutation is $1+2s$. Fisher [18] and Wright [90] derived a "transient distribution", $f(q,\gamma)$, to be the density of the frequency of a mutation,q,

$$f(q,\gamma) = \frac{1-e^{2\gamma(1-q)}}{1-e^{-2\gamma}} \times \frac{2}{q(1-q)} \tag{1.10}$$

For a particular site, suppose the frequency of a single mutation in the population is q, for a sample of n sequences, the probability of sampling i sequences of one

state and $n-i$ of another is binomially distributed with parameters n and q, denoted by $P(i|q)$. If mutations enter the population as a Poisson process with rate $\theta/2$, and assume sites evolve independently, then the number of sites that have i derived mutations in n sampled sequences, denoted by X_i, are independent Poisson distributed random variables with mean $\theta F(i,\gamma)$, where,

$$\begin{aligned} F(i,\gamma) &= \int_0^1 \frac{f(q,\gamma)}{2} \times P(i|q)dq \\ &= \int_0^1 \frac{1-e^{(-2\gamma(1-q)}}{1-e^{(-2\gamma)}} \times \frac{1}{q(1-q)} \begin{pmatrix} n \\ i \end{pmatrix} q^i(1-q)^{(n-i)}dq \end{aligned} \quad (1.11)$$

Sawyer and Hartl [71] showed that a random field composed of the individual frequencies of the derived mutations with expected density of $\frac{\theta}{2}f(q,\gamma)$ is a Poisson random field (PRF).

1.4 *Site Frequency Spectrum*

The site frequency spectrum (SFS) is the summary of the single nucleotide polymorphism (SNPs) frequency data for a genomic region in terms of a vector X such that X_i is the number of SNPs at frequency i out of n in the sample where n is the number of chromosomes sequenced, denoted by $X = [X_1, X_2, X_3, \ldots, X_{n-1}]$. A simple example of the SFS is illustrated in Table 1.1. For 6 sampled DNA sequences at the region of interest, we first aligned them with an out-group sequence at the same region to find the ancestor state for each polymorphic site under the infinite-sites assumptions. Then the number of mutations for those sites are easily identified, and the site frequency spectrum of this sample is $X = [2, 3, 0, 1, 1]$ since there are two sites that only one single mutation are observed, three sites has two

Table 1.1: Site Frequency Spectrum (SFS)

	SNP site						
	1	2	3	4	5	6	7
sequence 1	A	A	T	C	G	C	T
sequence 2	A	A	T	C	C	C	A
sequence 3	A	A	A	G	G	C	T
sequence 4	A	C	T	C	G	C	T
sequence 5	T	C	T	C	G	G	T
sequence 6	A	A	A	C	C	G	T
ancestor state	A	A	T	C	C	C	A
mutants number	1	2	2	1	4	2	5
SFS	$X = [2, 3, 0, 1, 1]$						

mutations, no site has three mutations, one site has four mutations and one site has five mutations.

In the Poisson random field (PRF), we assume X_i's are independently Poisson distributed with mean rate $\theta F(i, \gamma)$, then the probability of observing x_i sites that have i derived and $(n - i)$ ancestral mutations is

$$P(X_i = x_i|\theta, \gamma) = e^{-\theta F(i,\gamma)} \frac{(\theta F(i, \gamma))^{x_i}}{x_i!} \tag{1.12}$$

1.5 *Outline*

While we have introduced basic background knowledges relates to the projects, in the following chapters we will describe in details about the projects one by one.

In Chapter 2, I present a novel composite-likelihood-ratio test (CLRT) for detecting genes and genomic regions that are subject to recurrent natural selection (either positive or negative).

In Chapter 3, I introduce a new method for modeling unknown correlation among components of the SFS, called "Poisson Pairwise Difference Method (PPDM)" and show how this method can be applied to efficiently co-estimate the selection coefficient, γ, and mutation rate, θ, from arbitrarily correlated site frequency spectrum data.

Lastly, in Chapter 4, I develop a multiple linear regression model to estimate recombination rate variation from single nucleotide polymorphism (SNP) data. I demonstrate that this approach provides a novel, powerful and efficient way to detect hot spots for large genomic region.

Chapter 2

A Composite-Likelihood Approach for Detecting Directional Selection from Site Frequency Spectrum

2.1 *Background*

Among models for estimating mutation and selection parameters in various population genetic settings when DNA mutations are assumed unlinked, the Poisson random field (PRF) approach (Sawyer and Hartl 1992 [71]; Hartl *et al.* [24]; Wakeley 2003 [83]; Willamson 2003 [91]; Willamson *et al.* 2005 [93]) has proven quite useful. The inference rationale behind the approach is that natural selection will alter the site-frequency spectrum (SFS), making it possible to estimate the strength of selection needed to explain observed deviations from the neutral SFS expectations. However, distinguishing the effect of demographic history from that of natural selection can be very difficult. For example, patterns of neutral DNA variation linked to a site on which balancing selection is acting can be similar to sequence variation sampled from population with subdivision (Hudson 1990 [31]), and patterns of sequence under the effect of selective sweep are similar to those in an expanding population (Simonsen *et al.* 1995 [74]) or a recent bottleneck (Stephen *et al.* 1998 [77]; Galtier *et al.* 2000 [22]). Maximum likelihood method is a classical approach that is widely used in detecting natural selection and estimating population parameters (Griffiths and Marjoram 1996 [23]; Yang 1997 [94]; Nielsen 1998 [63]; Kuhner *et al.* 2000 [52]; Fearnhead and Donnelly 2001 [13]).

However, full likelihood inference methods use computationally intensive statistical techniques (either Markov chain Monte Carlo or importance sampling), with very substantial computational burdens and practically impossible for many data sets when the derivation of exact likelihoods is difficult (Rannala and Slatkin 2000 [68]). Composite likelihood has been frequently used recently to reduce the computational complexity so that it is possible to deal with large datasets and very complex models. It has good theoretical properties and it behaves well in many complex applications, for example, estimating recombination rate developed by Hudson (2001 [32]), subsequently adapted by McVean et al (2002 [58]) and further applied by Fearnhead and Donnelly (2002 [14]); detecting local signature of hitchhiking by Kim and Stephan (2002 [42]). In this project, we will derive a composite likelihood approach in the PRF using all of the information in the SFS regarding natural selection as opposed to traditional summary statistics of the data such as Tajima's (1989 [79]) D and the methods of Fay and Wu (2000 [12]) and Fu and Li (1993 [19]).

To model the effects of natural selection on the site-frequency spectrum, several assumptions are made within the standard PRF models (Sawyer and Hartl 1992 [71]; Hartl *et al.* 1994 [24]):

1. Panmictic population of constant size;

2. Weak selection with no dominance;

3. Equal selective effects of all nonlethal mutations;

4. Free recombination among segregating sites;

5. Infinite-sites mutation model;

6. No epistatic effect among mutations.

Wakeley (2003 [83]) has developed models that relax assumption 1. by considering an infinite-demes population structure; Williamson *et al.* (2004 [92]) have developed PRF models with dominance, relaxing assumption 2.; and Bustamante *et al.* (2003 [6]) and Sawyer *et al.* (2003 [72]) have modeled the effects of a distribution of selective effects among nonlethal mutations [relaxing assumption 3.]. The purpose of this project is to relax assumption 4. for the purpose of inference.

Since the PRF model assumes independence among sites, the application of the LRT for most genetic data is quite limited unless the assumptions of free recombination among sites can somehow be relaxed. One can imagine two potential solutions to the problem: (1) explicitly modeling natural selection and recombination to evaluate the true likelihood function via the ancestral selection graph (Krone and Neuhauser 1997 [50]; Neuhauser and Krone 1997 [62]; Slade 2001 [75]), and (2) taking a composite likelihood approach by continuing to treat sites as independent and then correcting parameter estimates and critical values for the LRT via simulation. From a statistical point of view, the former approach is more desireable, since the likelihood function contains all the information about natural selection available in the data (*e.g.*, distribution of haplotypes, patterns of linkage disequilibrium). Unfortunately, full likelihood inference is so computationally costly as to be out of reach for practical sample sizes at single loci and certainly out of reach for genome-wide analyses. Therefore, due to practical motivations, we investigate the composite likelihood approach here, since the composite likelihood solution for a single locus can easily scale to genome-wide levels and can be expanded to include increasingly realistic demographic scenarios.

In this project, we explore the performance of a composite likelihood ratio

test (CLRT) for recurrent directional selection under varying levels of selection, mutation and recombination while relaxing the assumption of independence among sites. The initial motivation for this project was Bustamante *et al.*'s (2001 [5])result that the LRT proposed by Hartl *et al.* (1994 [24]) is not robust to deviations from the assumption of independence among sites (*i.e.*, the test has a much higher type I error than expected). However, by modifying the critical value of the LRT statistics, a proper test conditional on an estimate of the population recombination rate could be constructed with desired size (type I error, we will use size and type I error interchangeably afterwards), we refer to this test as the composite likelihood ratio test (CLRT) to distinguish it from the LRT designed from independent data and to signify that we are *not* dealing with the true likelihood function of the data under recombination and selection, but rather an approximate likelihood function. If the data come from a population with the same demography we have used for our neutral simulations (*e.g.*, a panmictic population of constant size) and our estimate of the recombination rate is accurate, such a test would be guaranteed not to reject neutrality more often than expected (namely, $100 * \alpha\%$ of the time). One property of interest of this project is the distribution of the CLRT statistic and how it changes with the level of recombination rate ($R = 4N_e r$) (See table 2.1 for notations).

Given a data set, there are two main approaches that have been employed for estimating . One method is based on observing the frequency of sequence exchange between distant markers (Ashburner 1989 [2] ; True *et al.* 1996 [80]; Bouffard *et al.* 1997 [4]; Nagaraja *et al.* 1997 [60]), the other method estimates from the patterns of sequence variation expected in a random sample of DNA sequences from a population (Hudson and Kaplan 1985 [29]; Hudson 1987 [30]; Griffiths, R. and P.

Table 2.1: Notations used

N_e	Effective population size
r	Per-locus recombination rate per generation; $R = 4N_e r$
μ	Per-locus mutation rate per generation; $\theta = 4N_e \mu$
s	Relative fitness of the mutant; $\gamma = 2N_e s$
m	Migration rate (proportion of migrants in the subpopulation per generation); $M = 4N_e m$
D	Number of demes
β	Population growth rate
n	Number of sequences sampled
S	Total number of segregating sites in the sampled sequences
Q	Number of replicates in the simulation study
R_h	Recombination rate estimator from Hudson (1987 [30])
R_{hw}	Recombination rate estimator from Hey and Wakeley (1997 [25])
Λ	Test statistic of the CLRT
Λ^*	95% critical value of the CLRT
t_{bs}	Time in the unit of $4N_e$ generations ago that the bottleneck happens
t_{be}	Time in the unit of $4N_e$ generations ago that the bottleneck recovers from the bottleneck
f	Ratio of the population size during bottleneck to the original size

Marjoram 1996 [23]; Hey and Wakeley 1997 [25]; Wakeley 1997 [81]; Kuhner (et al.) 1999 [51]). Since here we are interested in the local recombination rate, we use the latter approach. Among them, Hudson (1987 [30]) proposed an estimator of recombination rate in a finite population model without selection. Hey and Wakeley (1997 [25]) derived the method of estimating recombination rate by coalescent theory based on multiple subsets consisting of four sequences. Simulation study (Wall 2000 [85]) shows that both Hudsons (R_h) and Hey and Wakeleys estimator (R_{hw}) perform well with large sample size ($e.g. n = 50$) and improve as the mutation rate increases. However, comparing these two estimators with all others, Wall (2000 [85]) shows that R_h over-estimates R (a large proportion of R_h/R greater than 5.0); while R_{hw} under-estimates R (with majority of R_{hw}/R less than 0.2). In this project, these two extreme estimators are explored. Our approach is that given the data, we apply Hudsons and Hey and Wakeleys methods to estimate the recombination rate, and for different levels of the recombination, we modify the critical value of the CLRT statistics such that the test will approach the desired size. In this way, we could apply the CLRT derived from PRF model to detect directional selection without the strict independent sites assumption while still controlling the type I error. A potential pitfall of such an approach is that there are several putative alternative hypotheses to a single null hypothesis. Populations where data were drawn from might not be panmictic or completely isolated and thus samples may contain some migrants which contribute to sequence polymorphism. For example, low levels of migration among subpopulations will elevate the proportion of observed high-frequency-derived mutations above their neutral expectation in a panmictic population much in the same way as positive selection (Nielsen 2001 [65]; see also Figure 2.1). Furthermore, natural populations usually

fluctuate in their sizes in the evolution which may significantly affect the pattern of genetic variations. Such as a sample of DNA sequences drawn from an exponentially growing population can look like a sample from a population of constant size subject to weak negative selection (*i.e.*, both scenarios lead to an excess of low-frequency variants *vis-á-vis* neutrality). Therefore, another issue of interest concerns the robustness of the CLRT to demographic and population structure. We also explore the probability of rejection neutrality if the data (drawn from forward simulation) is truly under selection with varying levels of recombination, which addresses the power of the CLRT (the probability of rejecting the null hypothesis when it is false) of neutrality without the assumption of independence among sites.

In the project, we (1) characterize the distribution of the CLRT statistic (Λ) as a function of the recombination rate (R); (2) explore the effects of the bias of different estimators of R on the size of the CLRT; (3) explore the robustness of the model to demographic factors and population structures(for example, population exponentially growing, bottleneck and migration); (4) explore the power of the composite likelihood test of neutrality under varying levels of mutation and selection as well as recombination; (5) explore the power of distinguishing negative selection from population growth; (6) evaluate the performance of maximum composite likelihood estimation of selection coefficient.

2.2 *Materials and Methods*

Let $\underline{X} = [X_1, X_2, ..., X_{n-1}]$ represent the site frequency spectrum for a genomic region of interest such that X_k is the number of sites along the sequence that have k derived mutations and $n-k$ ancestral mutations, where n is the number of sampled

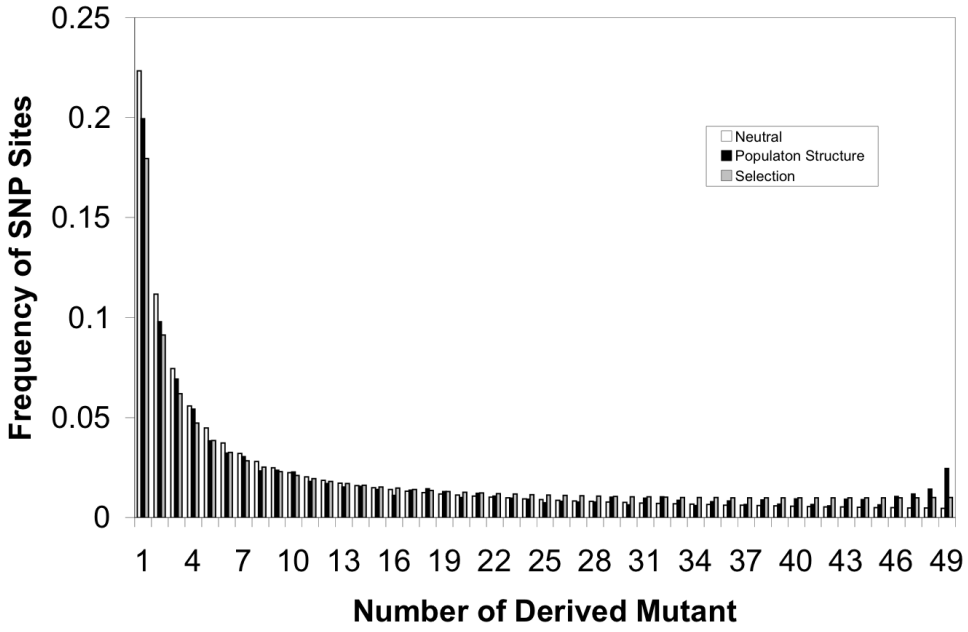


Figure 2.1: *Comparison of expected site frequency spectra for three scenarios. "Neutral" is the expected SFS under the standard neutral model (see Hudson 1990 [31]). "Population structure" is the expected site frequency spectrum for neutral mutations in a two-deme model with low symmetric migration rate ($4N_em = 0.2$) found via* 1000 *coalescent simulations using "ms" (Hudson 2002 [33]). "Selection" is the expected SFS under genic selection for the model described by Hartl* et al. *(1994 [24]). We use a value of $2N_es = 1.353$, which maximizes the likelihood of the expected population structure data under the selected model. As one can see, the site frequency spectrum under population structure can look similar to that under recurrent positive selection.*

sequences (throughout we assume the directionality of mutation is known, that is we know the ancestral state of the mutations). For γ defined as before, when $\gamma = 0$, the population is evolving neutrally; when $\gamma > 0$, it is under positive selection; and when $\gamma < 0$, it is subject to negative selection. Sawyer and Hartl (1992 [71]) showed that X_k is a Poisson distributed random variable with mean $\theta F(k,\gamma)$, where $F(k,\gamma)$ is in the form of the equation 1.11. Since under the model, the X_k's are independent, the likelihood function

$$L(\theta,\gamma|\underline{X}) = \prod_{i=1}^{n-1} \frac{e^{-\theta F(i,\gamma)}(\theta F(i,\gamma))^{x_i}}{x_i!} \tag{2.1}$$

Our null hypothesis is that the population evolves neutrally, $H_0 : \gamma = 0$, while the alternative hypothesis is the complement, $H_a : \gamma \neq 0$. The likelihood ratio test statistic proposed by Hartl *et al.* (1994 [24]) and investigated by Bustamante *et al.* (2001 [5]) for comparing these hypotheses is

$$\begin{aligned} \Lambda &= (-2\,log\,L(\hat{\theta}_w, 0|\underline{X})) - (2\,log\,L(\hat{\theta},\hat{\gamma}|\underline{X})) && (2.2)\\ &= 2\,(\sum_{i=1}^{n-1} x_i log\,F(i,\hat{\gamma}) - S\,log \sum_{i=1}^{n-1} F(i,\hat{\gamma}) + \sum_{i=1}^{n-1} x_i log\,i + S\,log \sum_{i=1}^{n-1} 1/i) && (2.3) \end{aligned}$$

where $\hat{\theta}_w$ is the maximum likelihood estimate (MLE) of θ under the neutrality, which turns out to be Ewens' (1974 [11]) and Watterson's (1975 [87]) estimator of θ; $\hat{\theta}$ and $\hat{\gamma}$ are the MLEs of θ and γ, respectively, under the full model with selection, found by maximizing the profile log-likelihood function as described in Bustamante *et al.* (2001 [5]), and $S = \sum_{i=1}^{n-1} x_i$ is the observed number of segregating sites. Under the assumption of independence among sites, Λ is asymptotically χ_1^2-distributed (Kendall 1987 [41]).

If sites are not evolving independently, the test statistic Λ does not necessary

following χ^2_1 distribution and the likelihood ratio test will have an unacceptably high type I error demonstrated by simulation studies (Bustamante *et al.* 2001 [5]). The reason for this is that the likelihood of the data in the presence of linkage is not simply the product of the likelihood across SNPs. That is, if sites are linked, Equation 2.1 is not the true likelihood function of the data, but rather a composite likelihood function, and the LRT statistic no longer corresponding to a true likelihood ratio test, but rather to a CLRT. Under such a scenario the distribution of the test statistic is no longer χ^2_1, but rather depends on the rate of recombination among sites. We must, therefore, use coalescent simulations with recombination to find the critical value Λ^* for the test statistic whenever we wish to analyze data with linkage among SNPs. While the LRT has been shown to have excellent power and $\hat{\theta}$ and $\hat{\gamma}$ have been shown to have little bias under the independence assumption (Bustamante *et al.* 2001 [5]), nothing is known about the statistical properties of the CLRT or the composite maximum likelihood estimates of θ and γ.

2.2.1 *Simulations*

To explore above issues, we simulated five different types of data (Hudson's 2002 [33] "ms" program was used for all coalescent simulations). The first type of data is neutral from a population of constant size. These data were used to explore how quickly the CLRT statistic Λ converges to a χ^2_1 distribution as a function of R and to compare the effect of different estimators of recombination rates on the realized size of the test. The second, third, and fourth types of data were neutral data from (a) a single subpopulation in an island model, (b) a panmictic population that had recently expanded in size, and (c) a panmictic population

that had undergone a single bottleneck. These data were used to explore the effect of these demographic factors on the type I error of the test. The fifth type of data was generated by the *f*orward *i*nfinite-sites *s*imulation wit*h* s*e*lection and *r*ecombination (*FISHER*) program written in *ANSI C* by *Lan Zhu*. *FISHER* was used to generate polymorphism data with recurrent selection and recombination under an infinite-sites model assuming constant population size. We ran *FISHER* with $10N_e$ generations of burn-in and replicate data sets sampled every $2N_e$ generations. These data were used to explore the power of the test under varying levels of mutation and selection, as well as recombination. Robustness simulations: To explore the null distribution of the CLRT statistic, we generated neutral data from a population of constant size for seven levels of recombination, $R \in \{0, 1, 5, 10, 50, 100, 1000\}$, using Hudson's (2002 [33]) "ms" program. For each of three sample sizes ($n = 10, 50, 100$), we simulated 1000 replicate data sets with a fixed number of segregating sites ($S = 100$) and constant recombination rate. For each replicate, we apply the CLRT and retain the test statistic Λ. The distribution of the CLRT statistic and the trend that Λ varies greatly with R are plotted in Figure 2.2 and 2.3, respectively.

2.2.2 *Algorithms*

The algorithms for calculating the CLRT is as follows:

Algorithm 1: Composite Likelihood Ratio Test (CLRT):

1. Given an observed site-frequency spectrum, X_{OBS}, estimate $\hat{\theta}$ and $\hat{\gamma}$ using the one dimensional optimization described in Bustamante *et al.* (2001 [5]), and calculate the CLRT statistic Λ_{OBS} via equation 2.2.

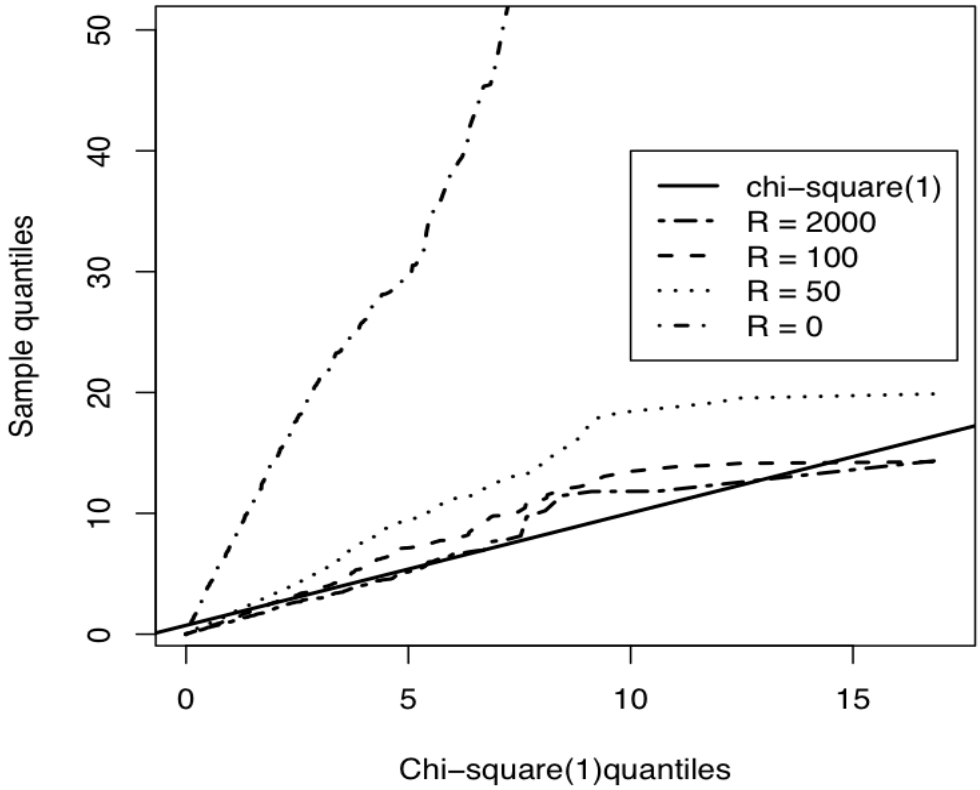


Figure 2.2: *Distribution of the test statistics (Λ) for the test assuming Hartl, Moriyama and Sawyer (1994 [24]) model as a function of population recombination rate (R). Y-axis is quantiles of Λ's calculated by CLRT from sampled sequences, X-axis is quantiles of data drawn from χ^2_1 distribution. Λ converges to χ^2_1 distribution with large R . 1000 replicates of data sets were sampled from Hudsons ms program, each with sample size $n = 50$, fixed number of segregating sites $S = 100$ and various level of recombination rate.*

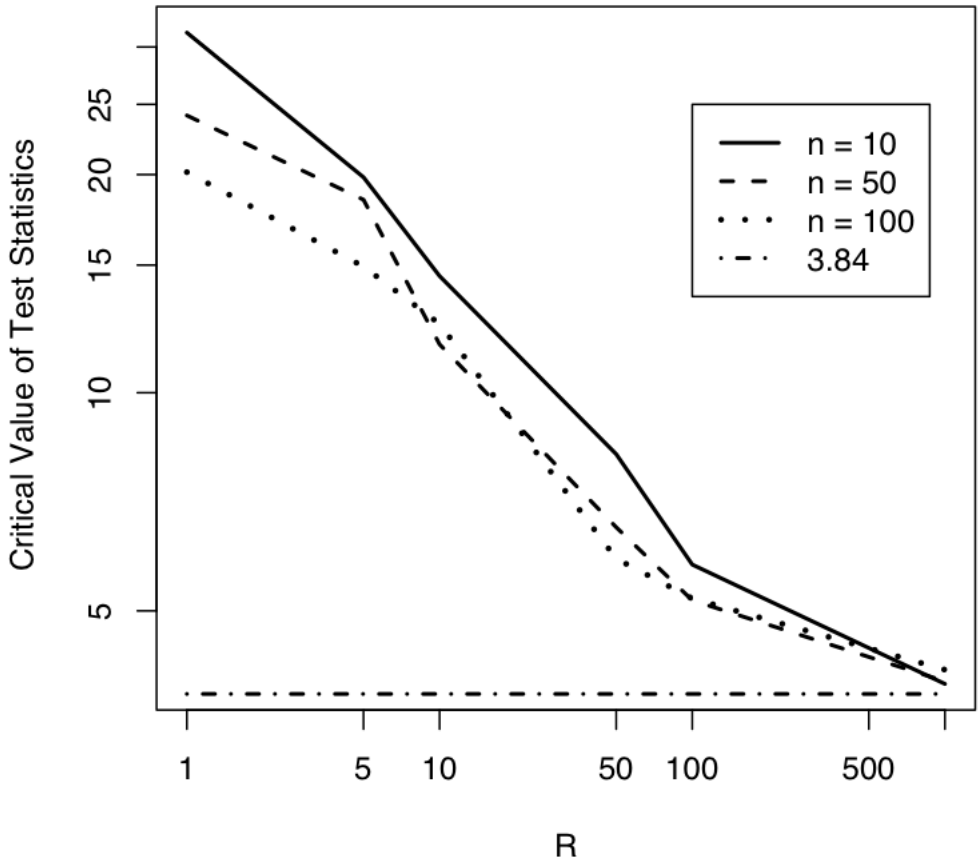


Figure 2.3: *95% critical value of the test statistic* (Λ^*) *converges to* $\chi^2_{1,0.95} = 3.84$ *(plotted in log scale for both* $x-$ *and* $y-$ *axes). Data were drawn from Hudson's (2002 [33])"ms" program with sample size* $n \in \{10, 50, 100\}$ *and fixed segregating sizes* $S = 100$.

2. Generate Q replicate data sets X_1, X_2, ..., X_Q from a standard neutral model with recombination rate R corresponding to the region of interest and S to the observed number of segregating sites in X_{OBS}. Apply the optimization in step 1 to each of the replicate data sets and generate the replicate CLRT statistics Λ_1, Λ_2, ..., Λ_Q.

3. The P-value for the CLRT corresponding to data X_{OBS} is estimated as $P(\Lambda_{OBS}|H_0) \approx (\sum_{i=1}^{Q} I(\Lambda_{OBS} \leq \Lambda_i))/Q$, where $I(\,)$ is the indicator function that evaluates to 1 if the argument is true and 0 otherwise.

In practice, the true recombination rate for sampled sequences is unknown and must be estimated from data. We were interested in investigating the effect of estimation bias in the recombination rate on the type I error of the CLRT. As Wall (2000 [85]) showed, there is no single best estimator of R and , in practice, most estimators do poorly if R is close to 0. Here we explored Hudson's (1987 [30]) and Hey and Wakeley's (1997 [25]) estimators since they tend to overestimate and underestimate R, respectively, for a broad range of values. Since Hudson's estimator has low reliability if data sets are not very large (Hudson 1987 [30]), we simulated data with sample size $n = 50$ and fixed segregating sites at $S = 100$. The detailed algorithm is as follows and the results of our analysis are summarized in Figure 2.4.

Algorithm 2: Estimating realized type I error of CLRT when R is estimated from data:

1. Generate neutral data X_{OBS} with known recombination rate R and apply the CLRT to obtain the test statistic Λ_{OBS}.

2. For X_{OBS} estimate R by Hudson's [30] and Hey and Wakeley's [25] methods

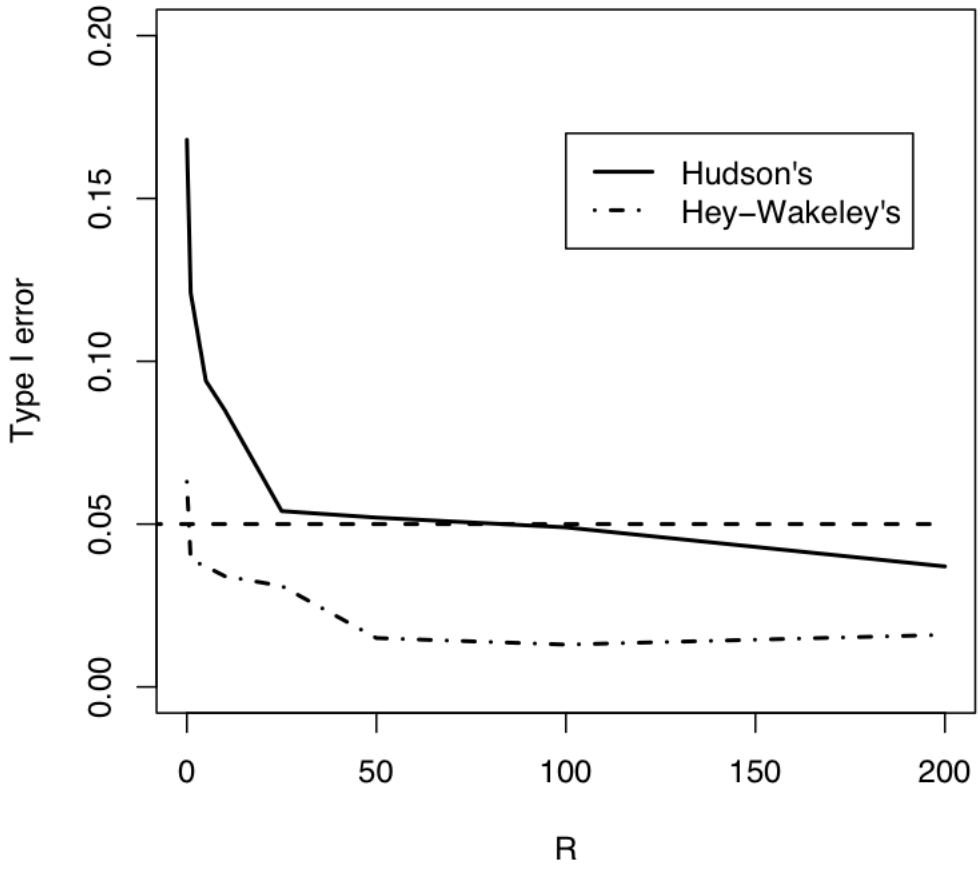


Figure 2.4: *Effect of the bias of the recombination rate estimator on the size of the CLRT. Data were drawn from Hudson's "ms" program [30] with sample size* $n = 50$ *and fixed segregating sites* $S = 100$*. Recombination rates were estimated by the "SITES" program (Hey and Wakeley 1997 [25]).*

using SITES [25] and denote the estimates R_h and R_{hw}.

3. Generate $Q = 1000$ replicate data sets with the same sample size and number of segregating sites as X_{OBS} under the estimated recombination rate, R_h. For each replicate, perform CLRT and keep test statistic Λ. The empirical $(1 - \alpha)$ quantile of the distribution of Λ among the 1000 replicates is the critical value of the test statistic Λ^* at α-level (for all simulations we used $\alpha = 0.05$). Similarly, we can find Λ^* with estimated recombination rate R_{hw}.

4. If $\Lambda_{OBS} > \Lambda^*$, reject the neutral hypothesis; otherwise, fail to reject at the $\alpha = 0.05$ level.

5. Repeat steps 1-4 1000 times. The proportion of the false rejection is the realized size of the CLRT under the PRF model when the recombination rate is not known.

In the current model, we assume no population structure to the data. We are interested in investigating how well the CLRT performs when this assumption is violated. We simulated the second type of data with sample size $n = 50$ and fixed number of segregating sites ($S = 100$), $R = 0$ under the island model for $D \in \{2, 5, 10, 20, 50\}$ (D is the number of demes), and $0 \leq M \leq 15$ ($M = 4N_e m$, where m is the fraction of each deme made up of new migrants each generation). The reason for fixing the number of segregating sites is that the distribution of the number of segregating sites changes with the migration rate if we fix the overall mutation rate of the entire population (Wakeley 2001 [82]). When we explore the effect of the migration rate on the size of the test, we want to control for the effect that is caused by the difference in the number of segregating sites. The detailed procedure is as follows and the results of this analysis are shown in Figure 2.5.

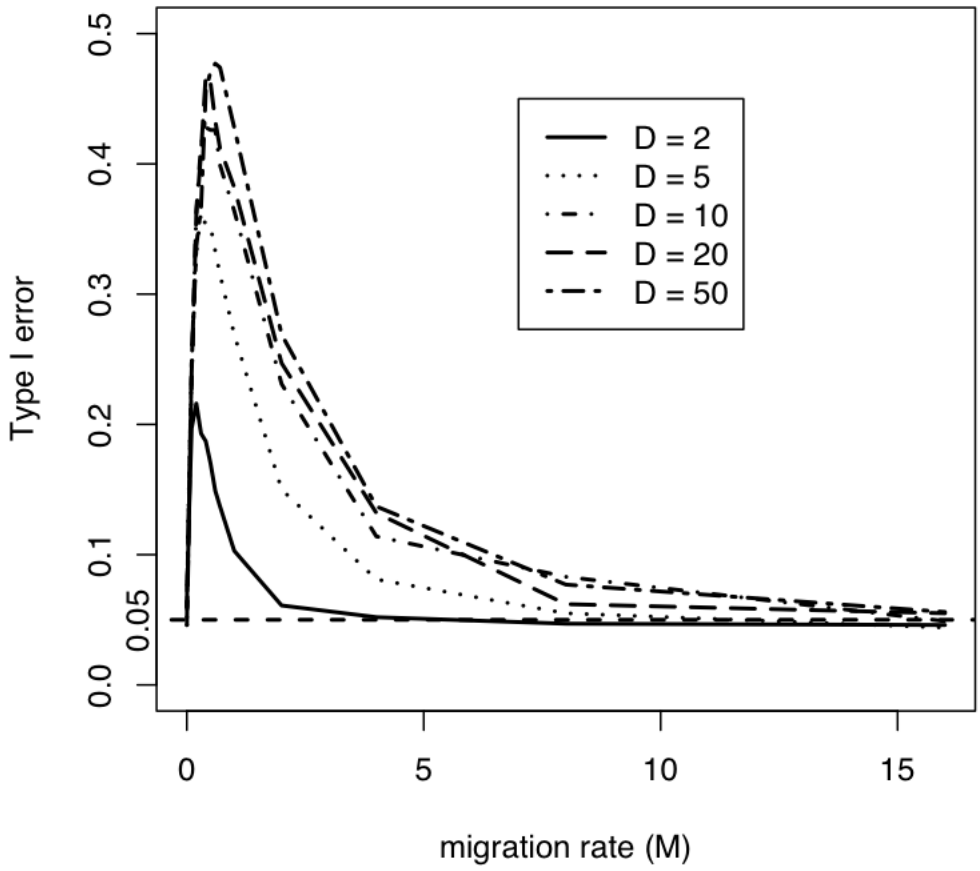


Figure 2.5: *Effect of population structure on the size of the CLRT. Data were drawn from the island model using Hudson's "ms" program [33] with given number of demes,* $D \in \{2, 5, 10, 20, 50\}$ *with* $R = 0$.

Algorithm 3: Procedure for estimating realized type I error of CLRT in the presence of population structure/ demographic history:

1. Generate $Q = 10,000$ data sets with S segregating sites from a panmictic population of constant size. Estimate the critical value of CLRT at the $\alpha = 0.05$ level as the $9501st$ largest value and denote this quantity as Λ^*.

2. Sample n sequences with $R = 0$ from a single deme out of D possible demes in the island model with migration for a given level of M. Apply the CLRT and retain the observed test statistic, Λ_{OBS}.

3. If $\Lambda_{OBS} > \Lambda^*$, reject the neutral hypothesis; otherwise, fail to reject at the $\alpha = 0.05$ level.

4. Repeat steps 2 and 3 1000 times for each parameter combination. The proportion of data sets that reject neutrality (*i.e.*, number of data sets out of 1000 with $\Lambda_{OBS} > \Lambda^*$) is the realized type I error of the CLRT.

Another assumption of the PRF model that may be problematic is the assumption of constant population size. To explore the effects of exponential growth [*i.e.*, the population size is given by $N(t) = N_e\, exp(-\beta t)$, where N_e is the present population size, t is the time before present, measured in units of $4N_e$ generations, and β is the growth rate], we modify step 2 of the above algorithm and generate data within "ms" for rates of growth $\beta \in \{0.1, 0.2, 0.4, 0.8, 1.6, 3.2\}$ (Figure 2.6). For a bottleneck, we simulate data for $n \in \{10, 100\}$ with various recombination rates $R \in \{0, 10, 100\}$, assuming the bottleneck happened at $t_{bs} = 0.025$ or 0.05 (in the unit of $4N_e$ generations) before the current sampling time and recovered to the current population size at $t_{be} = 0.0125$ (in the unit of $4N_e$ generations). We consider

two levels of the population reduction during the bottleneck, *i.e.*, $f \in \{0.1, 0.01\}$ (Figures 2.7 and 2.8).

Power simulations: To evaluate the power of the CLRT (*i.e.*, Probability of rejecting a false null hypothesis of neutrality), we wrote a forward simulation program, FISHER, to simulate a genomic region under recurrent selection and recombination using an infinite-sites model of mutation assuming constant population size. Power was estimated as the proportion of replicates generated under selection for which the null hypothesis was rejected by the CLRT. The detailed algorithm is as follows and results of the FISHER simulations can be found in Figures 2.9, 2.10, and 2.11.

Algorithm 4: Procedure for Estimating Power of CLRT:

1. Generate a data set of n sequences via FISHER given mutation rate, selection coefficient and recombination rate; apply CLRT to obtain the test statistic Λ_{OBS} and corresponding P-value from Algorithm 1.

2. If $P < 0.05$, reject the neutral hypothesis; otherwise, fail to reject at the $\alpha = 0.05$ level.

3. Repeat steps 1 and 2 500 times and calculate power of the test as the proportion of rejections.

The power of the test above is based on estimating the P-value assuming a constant population size. Since population growth may have similar effect as negative selection, we would like to examine how powerful the CLRT is in distinguishing negative selection from an exponentially growing population model. For these simulations, the data sets were simulated via FISHER given selection coefficient, mutation rate and recombination rate. The critical value of the CLRT was de-

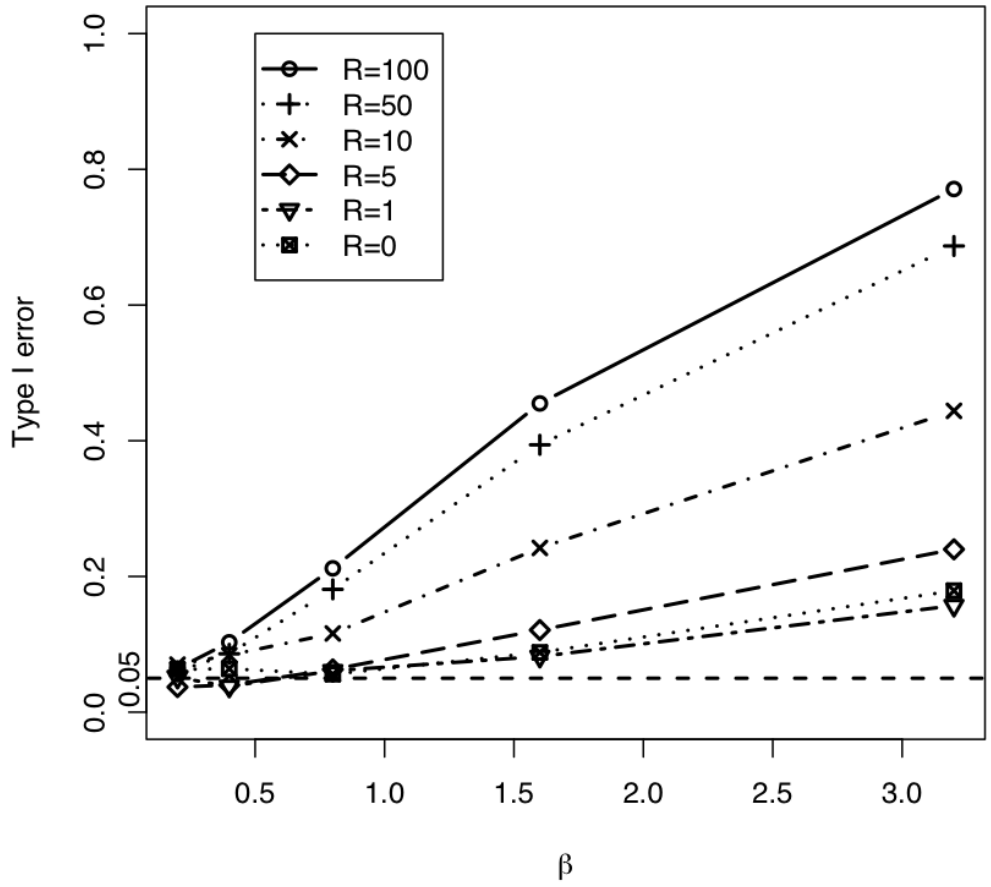


Figure 2.6: *Effect of the population size changes on the size of the CLRT. Data were drawn from the population exponentially growing model by Hudsons ms program [33] with sample size* $n = 50$*, fixed segregating sites* $S = 100$*, growth rate* $\beta \in \{0.1, 0.2, 0.4, 0.8, 1.6, 3.2\}$*, and various level of recombination rate.*

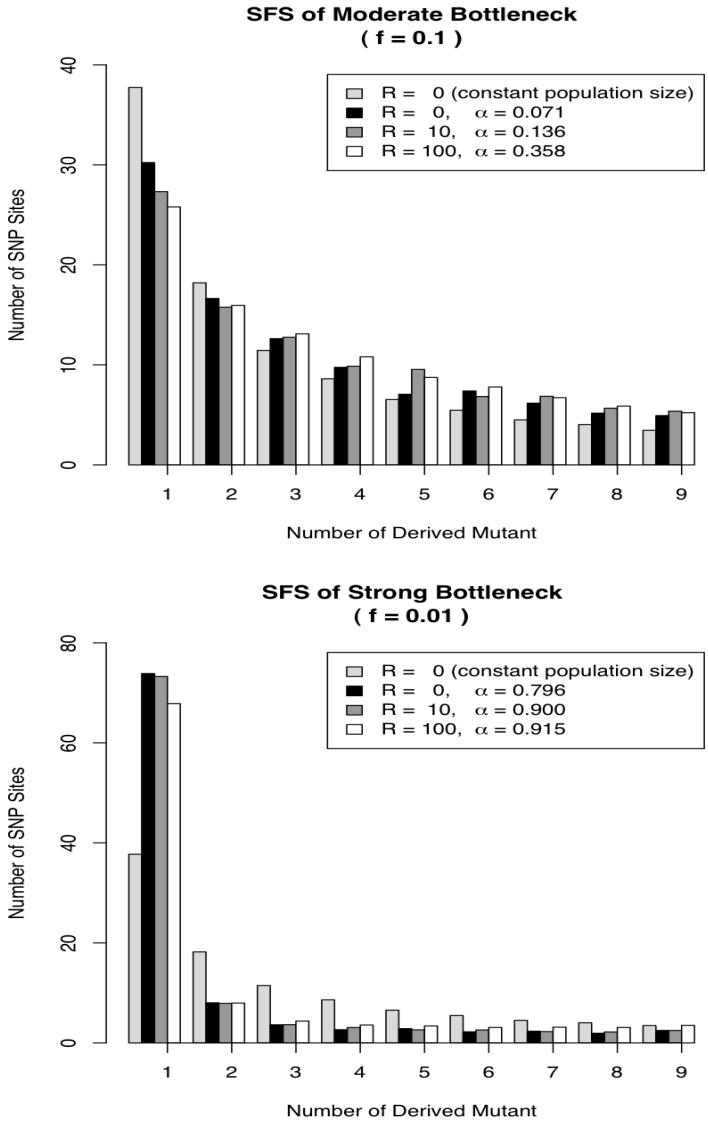


Figure 2.7: *Site frequency spectrum of data from a single population having undergone a recent bottleneck. Bottleneck occurred* $0.1N_e$ *generations ago, and it lasted* $0.05N_e$ *generations. Sample size* $n = 10$ *, with fixed segregating sites* $S = 100$*. f is the ratio of population size during bottleneck to the original size.* α *is the type I error of the CLRT. Top(A). Moderate Bottleneck with* $f = 0.1$*; Bottom(B). Strong Bottleneck with* $f = 0.01$*.*

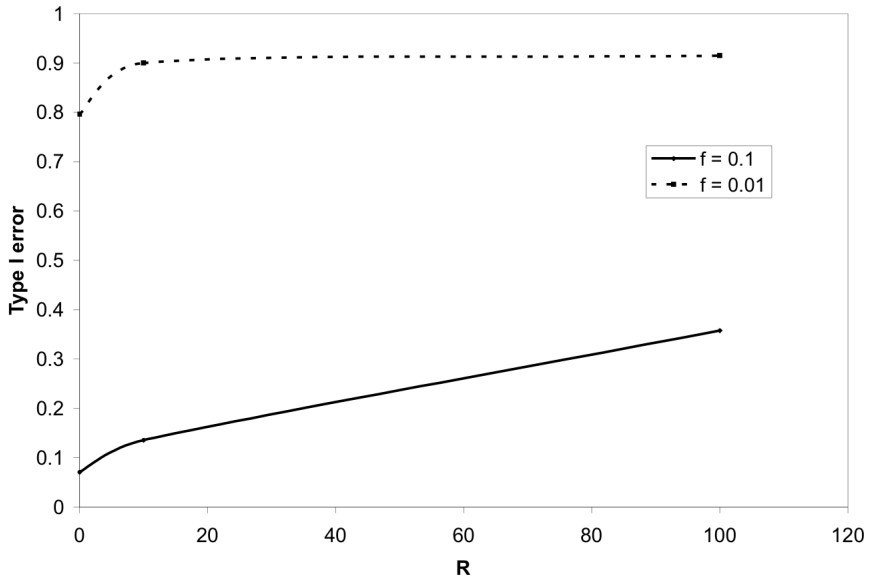


Figure 2.8: *Effect of recent population bottleneck on the size of the CLRT. f is the ratio of population size during bottleneck to the original size. Data sampling scheme is the same as that described in figure 2.7 A and B.*

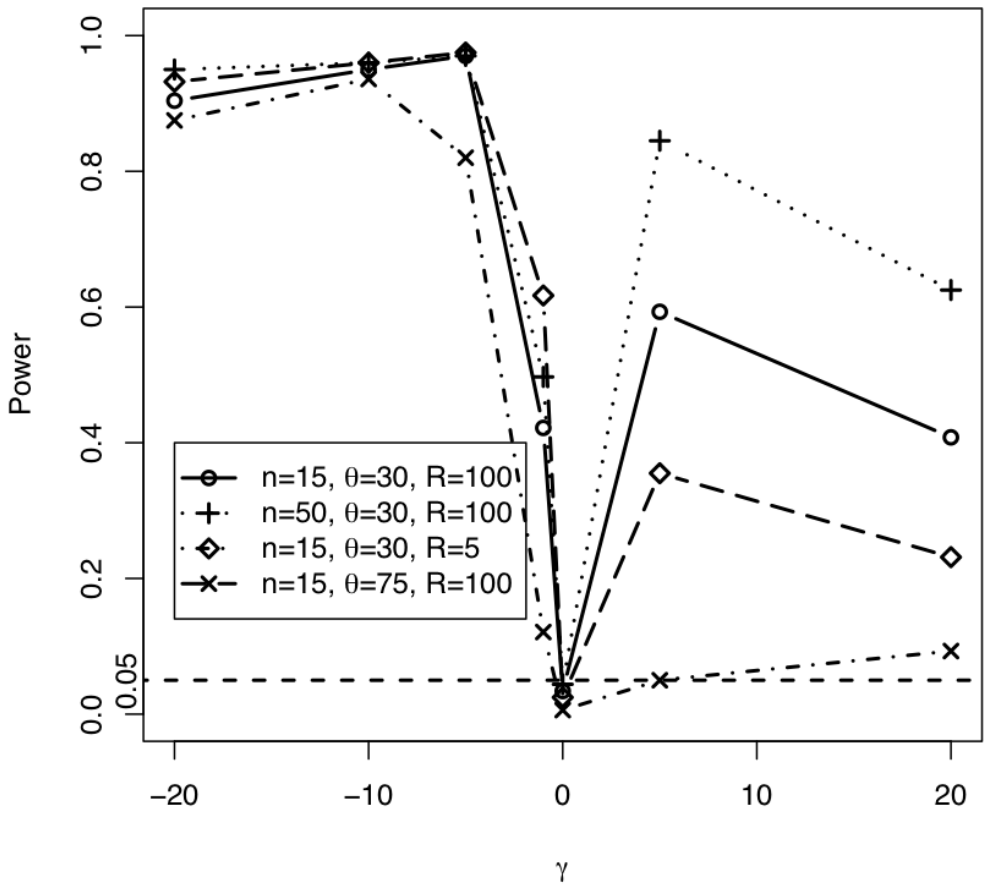


Figure 2.9: *Power of the CLRT under varying levels of selection. X-axis is the value of the selection parameter in the PRF model under which the data were simulated.*

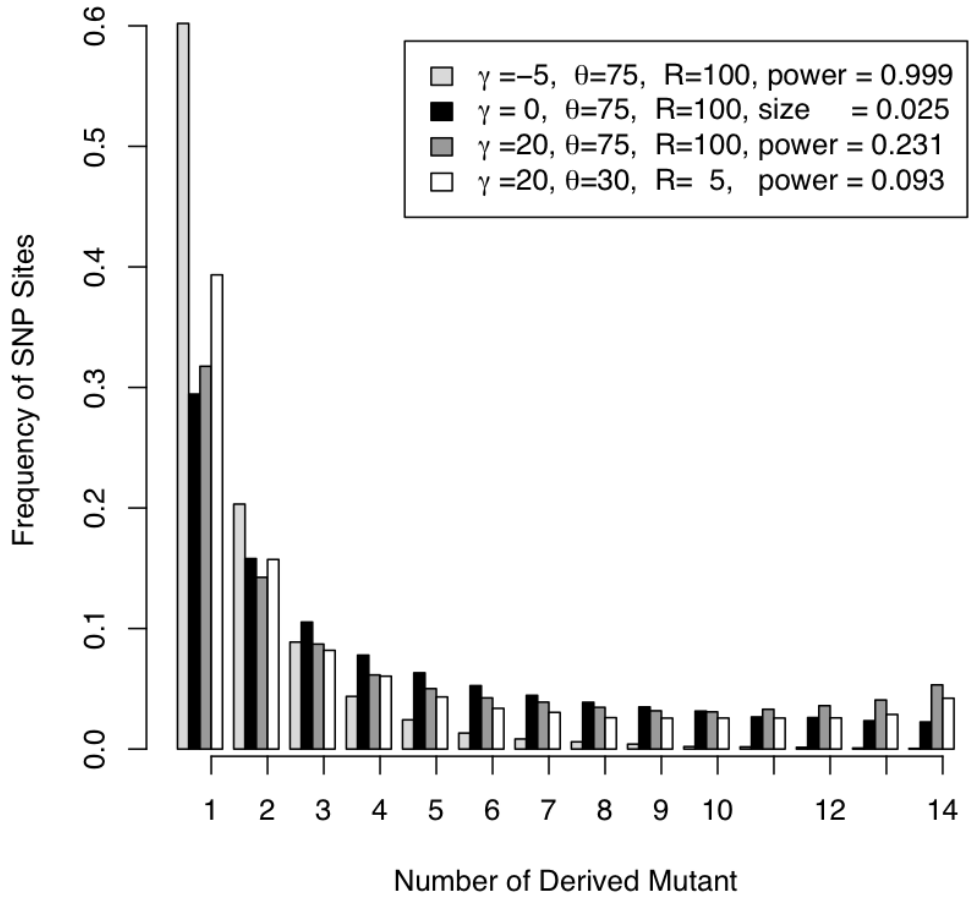


Figure 2.10: *Site frequency spectrum under recurrent negative selection, neutral and positive selection with varying levels of mutation and recombination rates. The Y-axis is the proportion of SNP sites that were found at frequencies* 1/15, 2/15,, 14/15.

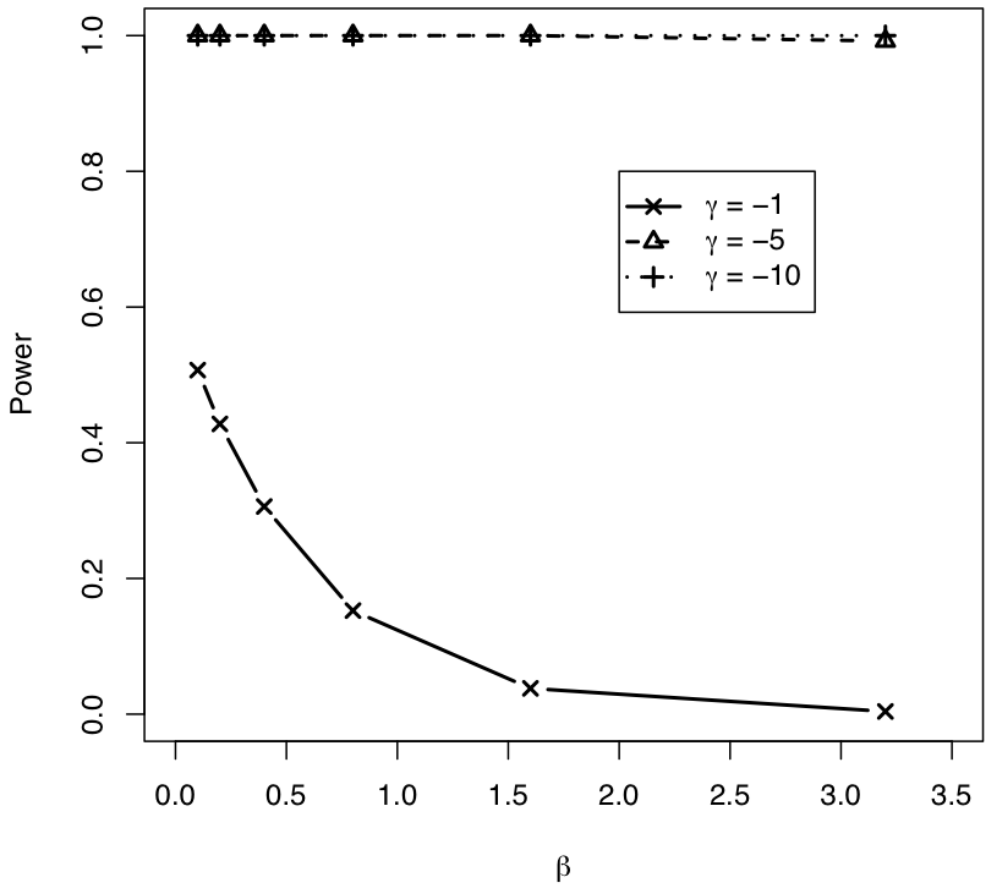


Figure 2.11: *Power of the CLRT in distinguishing negative selection from the population exponentially growing model. Data were simulated by FISHER program under the assumption of constant population size with sample size* $n = 50$, $\theta = 30$, $R = 100$ *under forward simulation model with selection coefficient* $\gamma = -1, -5, -10$, *respectively. X-axis is the growth rate* β, *the parameter of the data where the empirical distribution of the test statistics was obtained in order to get the critical value for the test.*

termined assuming the population has been growing exponentially, and mutations were neutral. We sampled 50 sequences with mutation parameter $\theta = 30$, recombination rate $R = 100$ and selection coefficient $\gamma \in \{-1, -5, -10\}$.We analyzed the power of the CLRT in distinguishing negative selection from the exponential growth with growth rate $\beta \in \{0.1, 0.2, 0.4, 0.8, 1.6, 3.2\}$. To achieve this, we modify step 2 in Algorithm 1 and simulate data under an exponential growth model. All other steps remain unchanged. (Results are shown in Figure 2.11.)

2.3 *Results and Discussion*

2.3.1 *How quickly does the test statistic (Λ) converge to a χ^2_1 distribution?*

From Figure 2.2, we confirm the theoretical prediction that the composite likelihood ratio test statistic Λ converges to a χ^2_1 distribution as recombination rate increases; unfortunately, the convergence rate is very slow. From Figure 2.3, we can see that the 95% critical value of Λ (denoted as Λ^*) does not attain the expected cutoff of $\chi^2_{1,0.95} = 3.84$ under the independence model until $R > 1000$ for all three levels of sample size considered ($n = 10, 50, 100$). Were we to test the neutral hypothesis using the CLRT and assume Λ followed a χ^2_1 distribution, the test would not attain the correct Type I error until the rate of recombination was inordinately large. This result is consistent with Bustamante *et al.* (2001 [5]) that the LRT is not robust to deviations from the assumption of independent among site, and highlights the need for developing a statistical method that can deal effectively with linkage among sites.

2.3.2 *How does bias in estimation of the recombination rate affect the realized size of the CLRT?*

We see from Figure 2.4 that the realized Type I error of the CLRT decreases with increasing recombination rate for both estimators studied. This is consistent with the fact that both Hey and Wakeley (1997 [25]) and Hudsons (1987 [30]) estimators improve as R increases (Wall 2000 [85]). In general, using R_h to estimate the recombination rate will result in larger Type I error than using R_{hw}. From our study, for $R \leq 15$, Hey-Wakeleys estimator performs better than Hudsons with size closer to the Type I error (0.05); for $15 \leq R \leq 125$, Hudsons method actually performs better than Hey and Wakely's, and for $R \geq 125$, both are overly conservative. Recalling that R_h is upwardly biased for low levels of R (Wall 2000 [85]), it becomes clear that overestimating the recombination rate leads to a lower Λ^* and hence an increased probability of rejecting the null hypothesis (and therefore larger Type I error). Consistent with this observation is that R_{hw}, which is downwardly biased, leads almost uniformly to a very conservative CLRT.

2.3.3 *How does undetected migration affect the size of the CLRT?*

Even if we had a perfect estimator of R, we might not attain a realized Type I error of $\alpha = 0.05$ due to other factors, such as population history. We see from Figure 2.5, that an island model of population subdivision is such a scenario. For all levels of D examined, the observe pattern is very similar: the Type I error of the CLRT is 0.05 at $M = 0$; it then increases sharply for $0 < M < 1$ and then decreases slowly to 0.05 as M increases towards infinity. In a structured population with

$M = 0$, all subpopulations are completely isolated and within each subpopulation, individuals undergo random mating. Since sequences subject to the CLRT are all sampled from one subpopulation assumed to be at equilibrium, it is not surprising to see that the realized size of the CLRT for data with $M = 0$ is at the proper level for all levels of D.

Slightly increasing the migration rate will impact the site frequency spectrum by reducing the relative proportion of low frequency SNPs and increasing the relative proportion of high frequency SNPs. This is due to the fact that if M is small but not too small, a sample of DNA sequences from a single subpopulation will often contain a single migrant from another deme. This migrant will, more often than not, be involved in the last coalescent event of the genealogy, since the rate of migration is small relative to the rate of coalescent for $M < 1$. This will cause an overrepresentation of gene genealogies that are stretched near the root and compressed near the external nodes. The site frequency spectrum, will thus, look similar to what is expected under positive selection, predicting an increase in the Type I error of the CLRT for neutrality.

As M gets larger, the proportion of a given subpopulation that originated in another deme increases linearly. And as M tends towards infinity, the fixation index F will tends to be zero [$\hat{F} = 1/(1 + 4N_e m)$, at equilibrium], indicating no population structure. Hence a sample of DNA sequences randomly drawn from a subpopulation would be wholly representative of the entire population and the CLRT should have Type I error at the desired level. Indeed, from our simulation study, when $M \geq 16$, for number of subpopulations < 10, all tests we studied have type I error ≤ 0.05. For large number of subpopulations (> 10), M should be > 32 in order to have proper size of the CLRT.

There are two possible ways to improve the CLRT *vis-á-vis* population structure.One is to modify the critical value of the CLRT by estimating M from neutral data and thus reducing the Type I error by producing a more sophisticated null model. The second approach is to jointly estimate selection and migration coefficients under various population structure models. It is important to note that both fixes might also introduce systematic bias in the realized Type I error due to bias in estimation of demographic parameters.

2.3.4 *How does recent population expansion affect the type I error?*

Another important assumption in the current model is the assumption of constant population size over generations. This assumption does not hold for the vast majority of species which we would like to analyze for evidence of natural selection at the genetic level. From Figure 2.6, we can say that CLRT is not robust against the assumption of constant population size though it does not do badly for relative tight linkage with low population growth rate. The Type I error increases with the population growth rate. The reason is that population growth causes an increase in the coalescent rate as the process proceed back in time, leading to star-like genealogies which results in an excess of mutations in external branches (*i.e.*, singletons or substitutions present in only one sampled sequence) (Tajima 1989 [79]; Slatkin and Hudson 1991 [76]). It is difficult to differentiate the site frequency spectrum of population growth data from that under negative selection. The larger the population growth rate, the more singletons and hence more likely to make false rejections. It is expected that recombination substantially affects the size of the CLRT which is shown to be true in figure 2.6. For small population

growth rate ($\beta < 0.1$), CLRT still performs very well with type I error ≤ 0.05 which means slight changes in the population size do not affect the size of the CLRT. Williamson *et al.* (2005 [93]) have recently developed a method that can jointly estimate selection and population growth assuming independence among sites. For that model, one can also perform the CLRT conditioning on the maximum likelihood estimate of the growth parameter from the rest of the genome and an estimate of the local recombination rate to simulate the critical value of the test statistic for a given gene.

2.3.5 *How does a recent population bottleneck affect Type I error?*

Simulation study reveals that the effect of population bottlenecks on the patterns of SFS is very complicated (Figure 2.7). Moderate bottlenecks (Figure 2.7A) result in less low-frequency SNPs and more medium- and high-frequency SNPs than under neutrality. Strong bottlenecks (Figure 2.7B) function in the opposite direction; namely, more rare SNPs than expected under the constant population size model. The reason for this is that rate of coalescence increases during the bottleneck period and depending on parameter values can look like either positive or negative selection (Galtier *et al.* 2000 [22]). For example, a recent weak bottleneck can lead to disproportionately longer internal branches as several lineages make it back into the ancestral population, and thus contribute to high frequency derived mutations which can look like positive selection. Alternatively, a very severe recent bottleneck will likely lead to the most recent common ancestor event during the bottleneck period, and thus to star-like external branches which may be difficult to distinguish from negative selection. As a consequence, the Type I error of the CLRT is quite

high in populations which have experienced a recent bottleneck event (Figure 2.8). Increasing sample size and mutation rates leads to even higher Type I error (results not shown).

While it is clear that the CLRT is not robust to the effects of a recent bottleneck, it may be possible to distinguish whether the rejection of the test is due to natural selection or the effect of the recent population bottlenecks. One approach is to use a composite likelihood Goodness-of-Fit statistic which measures concordance between the data and a selective model (Jensen *et al.*, 2005 [38]). Alternatively, the genomic distribution of CLRT statistic itself can be used, since a bottleneck would uniformly increase the proportion of loci across the genome that rejects neutrality.

2.3.6 *How powerful is the CLRT in detecting selection?*

To evaluate a statistical test, we not only want to control the Type I error, but also would like to assess the power [$1 - \Pr(\text{Type II error})$]. Our simulation results (Figure 2.9) suggest that CLRT has relatively good power to detect negative selection and moderate power to detect positive selection, if the population recombination rate is on the order of the mutation rate and there is moderately strong selection.

If natural selection is very weak ($|\gamma| < 1$) and sites are tightly linked, selection has little effect on the SFS and the CLRT, thus, has little power. When selection is strong and negative ($\gamma < -5$), the site-frequency spectrum is skewed towards rare alleles and the CLRT performs very well even for small sample size irrespective of the mutation or recombination rates. In detecting weak positive selection ($\gamma > 5$), the CLRT has medium power for moderate levels of recombination relative to mutation . We find that increasing sample size from $n = 15$ to $n = 50$ will uniformly increase power (Figure 2.9). However, increasing mutation rate from

$\theta = 30$ to $\theta = 75$, paradoxically, decreases the power for detecting positive selection. The statistical reason for this is that the site frequency spectrum of data with high mutation rate and tightly linked sites subject to weak positive selection is similar to the SFS from a neutral population (Figure 2.10). One biological reason for this phenomenon is that increasing the mutation rate (or reducing the recombination rate) increases interference among selected mutation, and thus reduces the overall efficacy of natural selection (Robertson 1961 [69]; Hill and Robertson 1966 [26]; Felsenstein 1974 [17]; Comeron and Kreitman 2002 [8]).

2.3.7 *Can the CLRT distinguish negative selection from the effect of population growth?*

As we see from Figure 2.11, the CLRT does not have much power in distinguishing very weak negative selection ($\gamma = -1$) from exponential growth. However, for moderately strong negative selection ($\gamma = -5$) , the CLRT has very high power to differentiate selection from exponential growth with growth rate in the range of $0.1 - 3.2$. This suggests the CLRT maybe particularly useful for finding genes that may be subject to moderate negative selection.

2.3.8 *Is MCLE (Maximum Composite Likelihood Estimator) a good estimator of selection coefficients?*

If the assumption of independence among sites is met, maximum likelihood estimation of the selection and mutation rate parameters performs very well (Bustamante *et al.* 2001 [5]). We are interested to know whether the estimator is still reliable when we relax the assumption of independence among sites. In Figure 2.12, we

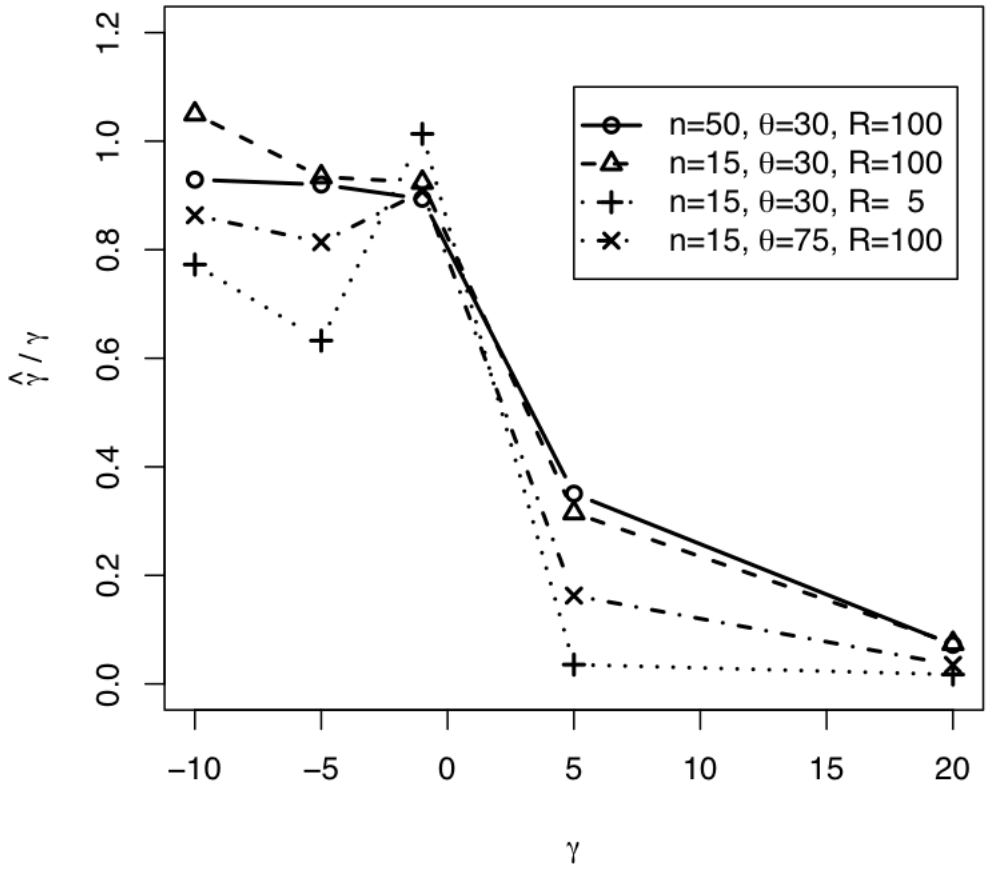


Figure 2.12: *$\hat{\gamma}/\gamma$ for data drawn from forward simulation with the recombination model (by the "FISHER" program). $\hat{\gamma}$ is the maximum likelihood estimator of the selection coefficient, and γ is the true parameter value under which the data were simulated.*

summarize the ratio of the MCLE of the selection parameter to the true selection coefficient as a function of γ. We can see that for weak negative selection ($\gamma \approx -1$), composite maximum likelihood estimation performs very well for all parameter combinations considered. Both mutation and recombination affect the accuracy of estimation. The parameter is under-estimated with higher mutation rate or less recombination events. In general, maximum composite likelihood estimator does not deviate far away from the true parameter value under which the data were simulated for negative selection with moderate mutation rate and have total recombination events > 100 per generation.

MCLE performs rather poorly in estimating the strength of positive selection in the presence of linkage with a large bias towards underestimation. The main reason is likely to be reduction in the effectiveness of selection because of interference among selected mutations (see Comeron and Kreitman 2002 [8]). That is, even if each mutation that enters the population has a selective advantage of, say , $N_e s = 5$, because there are few chromosomes that lack positively selected mutations, there will be only small fitness difference among chromosomes. As a result, mutations have a smaller realized effect on the site-frequency spectrum than predicted under the independence assumption.

2.4 *Conclusions*

The composite likelihood method presented here for inferring natural selection from DNA sequence data has reasonably good performance, in terms of power and robustness. One advantage over previous PRF tests is proper control of Type I error if DNA sites are linked. As expected, the method used to estimate the local recombination rate can have profound effects on the realized size of the test. We

predict this will be a general property of CL methods that aim to infer selection from standing patterns of genetic variation, and very little is known about the accuracy of methods for estimating recombination in the presence of recurrent selection.

We also find that undetected population structure, population growth, and/or bottlenecks can all inflate the realized Type I error of the test above its nominal level. One possible solution is to explicitly model selection and demography in future incarnations of the CLRT. In particular, by analyzing several unlinked loci simultaneously one may be able to estimate common shared parameters (such as expansion rate or time since bottleneck), while allowing for locus-specific selection parameters. Likewise, it is known that variation in selection among sites as well as dominance can have strong effects on the SFS (Bustamante *et al.*, 2003 [6]; Williamson *et al.*, 2004 [92].). We hope to incorporate these factors in future versions of the test.

Our simulation study shows that the composite likelihood ratio test has excellent power to detect negative selection and moderate power to detect positive selection. However, for weak selection $|\gamma| < 1$ and tight linkage $R < 5$, the method does not perform well, presumably due to interference selection. We have also shown that mutation rate and recombination rate profoundly influence the power of the CLRT.

It should be pointed out that a significant result of the CRLT (as with all test of neutrality) should be interpreted cautiously since there are several putative alternative hypotheses to single null hypothesis. Indeed, aside from the factors explored in this paper, processes such as population shrinking, inbreeding, and a single selective sweep, could also produce genealogies that are consistent with

some form of recurrent natural selection. Functional information will ultimately be needed to sort the false from true positives.

Chapter 3

The Poisson Pairwise Difference Method: A General Approach for Population Genetic Inference from SNP Data in the Presence of Correlated SNP Frequencies

3.1 *Introduction*

Single Nucleotide Polymorphism (SNPs) data are often used to make inference about the evolutionary history of population(s) or species from which they are sampled (reviewed in Morin *et al.* 2004 [57]). A particular pressing problem has been detecting the workings of natural selection from patterns of standing variation in natural populations with the goal of identifying mutations that may have been involved in molecular adaptation (*e.g.*, Lenormand *et al.* 1998 [53]; Barton 2000 [3]; Kohn *et al.* 2000 [48]; Schlotterer 2002 [73]; Nielsen *et al.* 2005 [66]). An important tool in this endeavor has been the analysis of the observed frequencies of segregating mutations or the site-frequency spectrum (SFS) (*i.e.*, the number of SNPs that are at a frequency $1/n$, $2/n$, , $(n-1)/n$ in a sample of n sequences, denoted by $X = [X_1, X_2, X_3, ..., X_{n-1}]$).

Many statistical tests of neutrality, for example, focus on comparing different summary statistics of the SFS to their expected distribution under a neutral model (Tajima 1989 [79]; Fu and Li 1993 [19]; Fu 1994 [20]; Fay and Wu 2000 [12]). The underlying rationale behind these tests is that since natural selection affects

components of the SFS differently, by identifying "skews" in the SFS, one may be able to characterize the strength and direction of natural selection. For example, mutations subject to weak negative selection will be found at lower frequencies than neutral mutations. Likewise, mutations experiencing balancing selection will tend to found at intermediate frequencies (Wayne and Simonsen 1998 [88]; Williamson *et al.* 2004 [92]) while positive (directional) selection, will results in a "U" shaped pattern of the SFS (*i.e.*, an relative excess of high- and low-frequency variants relative to mid-frequency variants.)

The tests of Tajima (1989 [79]), Fu and Li (1993 [19]), Fu (1994 [20]), Fay and Wu (2000 [12]) use only some part of the site-frequency spectrum and none of them take advantage of the complete information in the SFS. In recent years, the expected SFS has also been modeled under a myriad of demographic and selective scenarios under the assumption of independence among sites and infinite-sites mutation (Poisson Random Field models) including: genic selection in a population of constant size (Sawyer and Hartl 1992 [71]), general diploid selection in a population of constant size (Williamson, Alon, and Bustamante 2004 [92]), genic selection and population subdivision (Wakeley 2004 [84]), and genic selection and population size-change (Williamson et al. 2005 [93]). For each of these models, likelihood-ratio tests have been developed for comparing selective and neutral models, and for some of these models, we have a relatively good understanding of the statistical properties of the LRTs and corresponding maximum likelihood parameter estimates. For example, Bustamante et al (2001 [5]) explored the log-likelihood ratio test of directional selection proposed by Hartl et al. (1994 [24]) and demonstrated that test had very good power to detect directional selection when the ancestral states of all mutations in the sample are known (by aligning

with an outgroup), and that confidence intervals for the selection parameter in the model have the desired coverage.

One of the assumptions shared by all of the PRF models above is that given mutation rate and selection coefficient, each site is assumed to evolve independently according to an infinitely-many-sites mutation model (Sawyer and Hartl 1992 [71]). Population genetic theory, however, predicts that linkage will reduce the efficacy of selection (Robertson 1961[69]; Felsenstein 1974 [17]; Hill and Robertson 1966 [26]; Comeron and Kreitman 2002 [8]) and increase the variance of the components of the SFS (Watterson 1975 [87]; Hudson 1983 [27]; Fu 1995 [21]) potentially compromising the utility of the PRF approaches. In order to address this concern, Zhu and Bustanmante (2005 [95]) proposed a composite PRF likelihood ratio test (CLRT) which can be used in the presence of linkage among SNPs by modifying the critical value of the test statistic via coalescent simulations with limited recombination (Hudson 2002 [33]). We also demonstrated that the composite likelihood ratio test of neutrality in the modified PRF framework has excellent power to detect weak negative selection and moderate power to detect positive selection in the presence of linkage among selected mutations. An aspect of linkage which was not addressed in our previous work, however, is the potential for correlation among components in the site-frequency spectrum due to the underlying coalescent history of the sample.

Fu (1995 [21]) explored the statistical properties of segregating sites and derived the explicit formula for calculating the expectation and variance of each component of the SFS as well as the covariance between them under the standard neutral model with no recombination. He demonstrated, for example, that among all the variances and covariances, the covariance between X_i and X_{n-i} has the maximum

magnitude and is always positive. All other covariances are relatively small compared to the variances. The intuitive rationale behind this observation is that the last coalescent event in the genealogy joins two sub-trees: one containing i and the other $n-i$ lineages. Therefore, the length of this internal branch is positively correlated with X_i and X_{n-i} (since the longer the time the more mutations which can accrue that split the sample into i and $n-i$ lineages).

In this paper, we use Fus results as a starting point for addressing the issue of detecting selection from correlated site-frequency spectrum data within the PRF framework. We develop a new approach for modeling unknown correlation among components of the SFS, namely, Poisson Pairwise Difference Method (PPDM). The ultimate goal is to detect directional selection from arbitrarily correlated SFS data. We first demonstrate that for many demographic, structural and selective models, the only components of the SFS that show significant correlation with/without linkage are X_i and X_{n-i}. Next we explore the performance of Zhu and Bustamantes (2005 [95]) maximum composite likelihood estimation (MCLE) of mutation and selection parameters for both independent and correlated SFS. We then develop in detail the PPDM model for arbitrarily correlated SFS where the log-likelihood function is based on both selection and mutation parameters. We discuss how to find the maximum likelihood estimates (MLEs) of selection coefficient and mutation rate using Downhill simplex optimization in multi-dimensional parameter space (Nelder and Mead 1965 [61]) from multiple starting points. We also evaluate the performance of the PPDM in estimating selection and mutation parameters when the new vector Z, where $Z_i = X_i - X_{n-i}$ is used for inference. Likewise, we explore the power (the probability of rejecting the null hypothesis when it is actually false) of the log-likelihood ratio test of neutrality based on the PPDM model.

Lastly, we extend the PPD methodology and propose how it can be applied to infer other population parameters, for example, population growth rate.

3.2 *Theory*

3.2.1 *Poisson Difference Distribution*

Assume W_i for $i = 0, 1, 2$ are independently distributed Poisson random variables with parameter θ_i, respectively. Let $X = W_0 + W_1$ and $Y = W_0 + W_2$, then by the property of Poisson distribution, X and Y are also Poisson distributed with parameters $\theta_0 + \theta_1$ and $\theta_0 + \theta_2$, respectively with covariance θ_0. The joint distribution of X and Y is called bivariate Poisson distribution (Kocherlakota and Kocherlakota 1992 [47]) with the density

$$P(X = x, Y = y|\theta_0, \theta_1, \theta_2) = e^{(\theta_0+\theta_1+\theta_2)} \frac{\theta_1^x}{x!} \frac{\theta_2^y}{y!} \sum_{i=0}^{min(x,y)} \begin{pmatrix} x \\ i \end{pmatrix} \begin{pmatrix} y \\ i \end{pmatrix} i! (\frac{\theta_0}{\theta_1\theta_2})^i \tag{3.1}$$

In practice, the true covariance between the two variables X and Y may not be known, and if one is only interested in estimating θ_1 and θ_2, the model above may not be the ideal model with which to work.

To estimate θ_1 and θ_2 without estimating the covariance coefficient θ_0, we take advantage of the distributional property of the difference of two Poisson random variables and define a new random variable Z, where

$Z = X - Y = (W_0 + W_1) - (W_0 + W_2) = W_1 - W_2$

The distribution of Z is called Poisson difference distribution (Johnson et al.

1992 [39]) with density function as in 3.2

$$f(Z|\theta_1,\theta_2) = e^{(\theta_1+\theta_2)}(\frac{\theta_1}{\theta_2})^{z/2} I_{|z|}(2\sqrt{\theta_1\theta_2}) \tag{3.2}$$

where, $I_{|z|}(x)$ is the *Modified Bessel Function of order z as in 3.3*, (see Abramowitz and Stegum 1974 [1]) defined by

$$I_z(x) = (\frac{x}{2})^z \sum_{k=0}^{\infty} \frac{(\frac{x^2}{4})^k}{k!(z+k+1)} \tag{3.3}$$

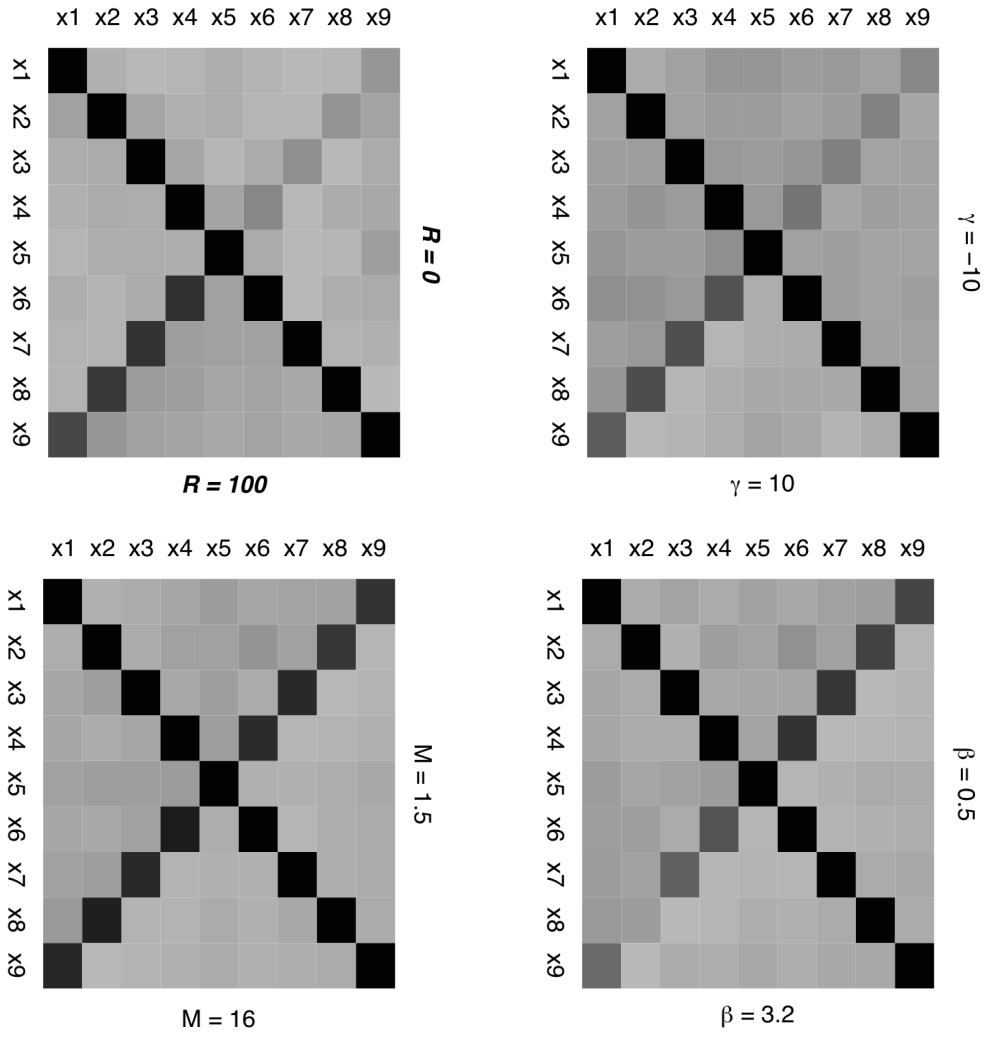
3.2.2 *Poisson Difference Distribution Applied on Site Frequency Spectrum*

If we assume no correlation among SFS components, under the assumptions of the PRF model (Sawyer and Hartl 1992 [71]), X_i's are independent Poisson distributed random variables with mean, $\theta F(i,\gamma)$, as in 1.11.

Below (figure 3.2.2) we show that the variance-covariance matrix of SFS components simulated under various demographic and selective scenarios suggests that the correlation of X_i and X_{n-i} has magnitude on the order of the variance of X_i, and probably ought not to be ignored. (We also demonstrate that the correlations between X_i and X_j for $j \neq i$ or $n-i$ are small in comparison (i.e., at least an order of magnitude smaller) and can thus be approximated as independent of each other). This is consistent with the derivations that Fu (1995 [21]) showed. Therefore, here we modify the PRF model and take the correlation of X_i and X_{n-i} into account.

Define $\theta^2\sigma_{i,j}$ as the covariance between X_i and X_j (notation consistent with that in Fu (1995 [21]). For $j \neq n-i$, we assume X_i and X_j are independent Poisson-distributed variables, and hence $\sigma_{i,j} = 0$. For $j = n-i$, $\sigma_{i,j} \neq 0$, we

Figure 3.1: *Heat map of correlations among the components of SFS. The darkness of the square indicates the strength of the correlation (the higher the correlation, the darker the square). The diagonals are the correlation between* X_i *and* X_{n-i}*. SFS were simulated with sample size* $n = 10$*,* $\theta = 30$ *under Hudsons ms or Zhu and Bustamantes FISHER program. Upper Left (A): From ms program with constant population size, recombination rate* $R = 0$ *(upper right) and* $R = 100$ *(lower left), respectively; Upper Right (B): Constant population size and no recombination, from FISHER program with selection coefficient* $\gamma = -10$*(upper right) and* $\gamma = 10$ *(lower left), respectively; Bottom Left (C): From ms program with migration, number of demes* $D = 5$*, all 10 sampled sequences were from one single deme with migration rate* $M = 1.5$ *(upper right) and* $M = 16$ *(lower left), respectively; Bottom Right (D): From ms program with population exponentially growthing, growth rate* $\beta = 0.5$ *(upper right) and* $\beta = 3.2$ *(lower left), respectively.*


x1 x2 x3 x4 x5 x6 x7 x8 x9
x1 x2 x3 x4 x5 x6 x7 x8 x9
R = 0
R = 100
x1 x2 x3 x4 x5 x6 x7 x8 x9
x1 x2 x3 x4 x5 x6 x7 x8 x9
γ = −10
γ = 10
x1 x2 x3 x4 x5 x6 x7 x8 x9
x1 x2 x3 x4 x5 x6 x7 x8 x9
M = 1.5
M = 16
x1 x2 x3 x4 x5 x6 x7 x8 x9
x1 x2 x3 x4 x5 x6 x7 x8 x9
β = 0.5
β = 3.2

rewrite it as $\sigma_{i,n-i}$. Under the modified PRF model assumptions, we assume X_i's are Poisson distributed with mean $\theta F(i,\gamma)+\theta^2\sigma_{i,n-i}$. Define a new random variable Z_i, $Z_i = X_i - X_{n-i}$, $i = 1, 2, ..., m$. If n is odd, $m = \frac{n-1}{2}$, and if n is even, $m = \frac{n}{2}$. By the theories introduced above, Z_i (for $i \neq \frac{n}{2}$) follows Poisson difference distribution with the following density,

$$f(Z_i|\theta,\gamma) = e^{(\theta F(i,\gamma)+\theta F(n-i,\gamma))}(\frac{F(i,\gamma)}{F(n-i,\gamma)})^{z_i/2} I_{|z_i|}(2\theta\sqrt{F(i,\gamma)F(n-i,\gamma)}) \quad (3.4)$$

If $i = \frac{n}{2}$, $Z_i \sim Poisson(\theta F(\frac{n}{2},\gamma))$, with density:

$$f(Z_i|\theta,\gamma) = e^{\theta F(\frac{n}{2},\gamma)}\frac{(\theta F(\frac{n}{2},\gamma))^{z_i}}{z_i!}$$

For any observed SFS, we can obtain a corresponding Z vector, which is called Poisson Pairwise Difference Site Frequency Spectrum (PPDSFS). The expected PPDSFS in the model is:

$$\begin{aligned} E[Z_i|\gamma] &= E[X_i - X_{n-i}] \\ &= E[X_i] - E[X_{n-i}] \\ &= \theta F(i,\gamma) + \theta^2\sigma_{i,n-i} - \theta F(n-i,\gamma) - \theta^2\sigma_{i,n-i} \\ &= \theta[F(i,\gamma) - F(n-i,\gamma)] \end{aligned} \quad (3.5)$$

It can also be proved numerically that $E[Z_i|\gamma] = E[Z_i| - \gamma]$ for $i \neq \frac{n}{2}$ (see appendix), which unfortunately means that for odd sampled haploid sequences, applying this approach will not tell you the direction of the natural selection. However, for even sequences, due to the difference between $E[Z_{n/2}|\gamma]$ and $E[Z_{n/2}|-\gamma]$, the key component $Z_{n/2}$ plays an important role in this model and the PPD

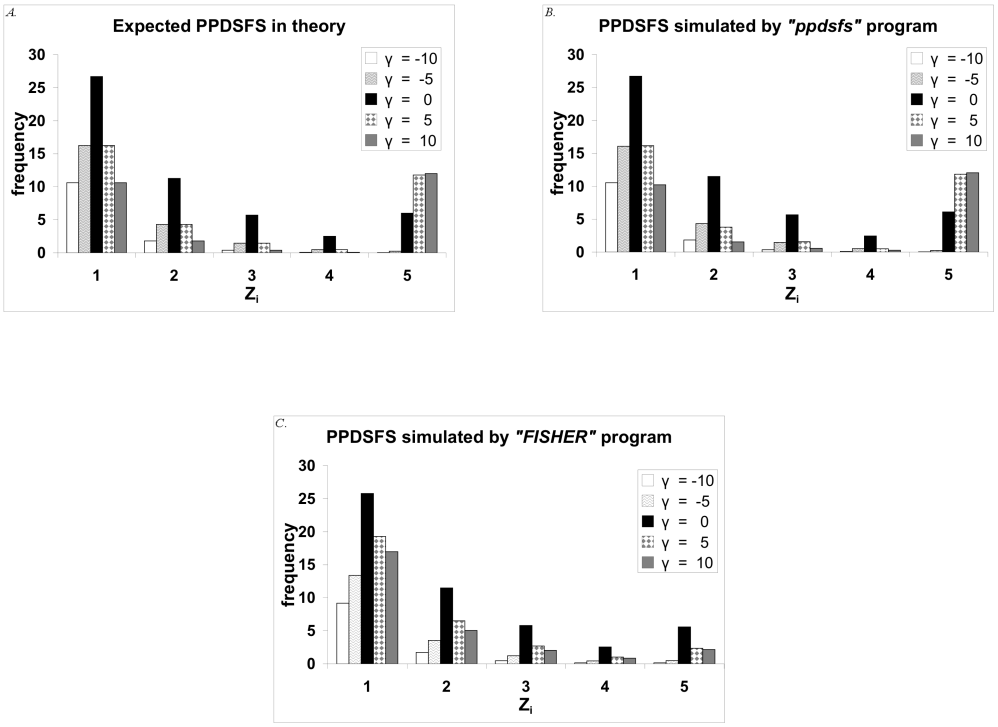


Figure 3.2: *The Poisson Pairwise Difference Site Frequency Spectrum (PPDSFS) under different selection pressure* $\gamma \in \{-10, -5, 0, 5, 10\}$. *A. Expected PPDSFS by coalescent theory; B. Data were simulated by ppdsfs program with sample size* $n = 10$, $\theta = 30$, $Z_i = X_i - X_{n-i}$, *for* $i = 1, 2, , [n/2]$. *C. Data were simulated by FISHER program with the same parameters as in B. The horizontal axis is* Z_i *and the vertical axis is the frequency of sites that has corresponding* Z_i *value from the SFS.*

method works well as we show below. An example of the expected and simulated PPDSFS under different selection pressure is shown in Figure 3.2.

For even sample size, the PPD model outline above leads directly to a likelihood-ratio test of neutrality, which compares the null hypothesis $\gamma = 0$ with the alternative hypothesis that $\gamma \neq 0$. Since we treat the Z_is as independent, the likelihood function is the product of the individual $f(Z_i|\theta,\gamma)$. Let $l(\theta,\gamma|Z)$ be the log-likelihood function for the PPD model,

$$\begin{aligned} l(\theta,\gamma|Z) &= \sum_{i=1}^{m} -\theta[F(i,\gamma)+F(n-i,\gamma)] \\ &+ \frac{z_i}{2} log\frac{F(i,\gamma)}{F(n-i,\gamma)} + logI_{|z_i|}(2\theta\sqrt{F(i,\gamma)F(n-i,\gamma)}) \\ &+ (1-(nmod2))(-\theta F(\frac{n}{2},\gamma)+Z_m log(\theta F(\frac{n}{2},\gamma))) \end{aligned} \tag{3.6}$$

3.2.3 *Maximum-log-profile-likelihood estimation*

To perform the likelihood-ratio test, we need to maximize (3.6) for θ and γ under the unconstrained parameter space and the constraint that $\gamma = 0$. Letting $\hat{\theta}$ and $\hat{\gamma}$ be the unconstrained MLEs of θ and γ, and $\hat{\theta_0}$ the MLE of θ under neutrality, the log-likelihood ratio test statistic is $\Lambda = l(\hat{\theta}),\hat{\gamma}|Z) - l(\hat{\theta_0},0|Z)$. We use Downhill Simplex method (Nelder and Mead 1965 [61]) in two-dimension from multiple starting points to find the $\hat{\theta}$ and $\hat{\gamma}$ for a given PPDSFS.

3.2.4 *Algorithm of LRT in the PPD model*

For an observed SFS X_{obs} of a genomic region, we can perform the following LRT in the PPD model:

1. Calculate PPDSFS vector Z_{obs} from the SFS vector X_{obs};

2. Apply Poisson Pairwise Difference Method (PPDM) on Z_{obs} to get the estimated MLEs $\hat{\theta}$, $\hat{\gamma}$ and the test statistic Λ_{obs};

3. Simulate X_1, X_2, ..., X_Q replicates of SFS from $ppdsfs$ program under neutrality ($ppdsfs$ is a program that is used to simulate SFS under the assumptions of PPD model with given mutation rate and selection coefficient) with the same sample size in step 1 and $\hat{\theta}$ in step 2. For each replicate of the SFS, apply PPDM to get $\hat{\theta}_i$, $\hat{\gamma}_i$ and the test statistic Λ_i;

4. The P value of the statistical test is $P = \frac{\sum_{i=1}^{Q} (\Lambda_{obs} \leq \Lambda_i)}{Q}$

3.2.5 *Properties of the LRT*

To measure the performance of the LRT for the PPD model, we evaluate the size of the test as well as the power of the LRT. The size of the test α is the probability of rejecting the neutrality when it is in fact true. To calculate the size of the test, we simulate 1000 SFS vectors X_{obs} s by $ppdsfs$ program under neutrality with known θ, and apply the LRT procedure described above for each replicate of the SFS; the proportion of false rejection of neutrality out of 1000 is the realized size of the test. The power of the LRT is the probability of rejecting the neutrality hypothesis when it is in fact false. Here we simulate a SFS vector X_{obs} by $ppdsfs$ program with known θ and γ, and apply the LRT described above. Repeating this procedure 1000 times, the power of the LRT is the proportion of the rejection (with P value less than 0.05) out of 1000 replicates. We would also like to see how PPDM LRT performs on the data from Zhu and Bustanmante (2005 [95])s FISHER program which is based on the forward simulation with selection and recombination algorithm. The detailed procedure of calculating the power of the

LRT is similar as described above by first simulating a SFS vector X_{obs} by FISHER with known θ and γ, then applying the modified LRT. The P value is calculated by simulating X_1, X_2,, X_Q replicates of SFS from Hudsons (2002 [33]) ms program under neutrality in step 3. All other steps remain the same.

3.3 *Results and Discussion*

3.3.1 *Correlation among components of the SFS*

To explore the correlation among components of the SFS, we simulated 1000 replicates of sequences with sample size 10 under Hudsons (2002 [33]) ms program (neutrality with constant population size, population exponentially growing and island migration model scenarios) and Zhu and Bustamantes (2005 [95]) FISHER program (with forward directional selection scenario), respectively. Figure 3.2.2 shows representative examples of the correlations among components of the SFS via the use of heat maps. In each heat map, the darkness of the square indicates the strength of the correlation. We can see from figure 3.2.2 that for all situations we explored, either with demographic, population structure scenario or under directional selection, the only highly or moderately correlated pairs are X_i and X_{n-i}(diagonals). All other pairs have correlation close to zero; therefore in our model those pairs are assumed to be independent of each other. Recombination will shuffle the diversity in the population, therefore reduces the variance or correlation between pairs of X_is (figure 3.2.2A). In an expanded population, where sampled lineages are more independent than those from small size population, we expect to observe less variance in the SFS with higher growth rate (figure 3.2.2B). Samples from one single population that undergoes low rate migration may have

migrants that most likely are the one branch that leads to the most recent common ancestor and it takes longer time to coalesce than that in the no population structure case. Therefore, relatively longer T_2 branch results in higher correlation between X_i and X_{n-i}(figure 3.2.2C). Directional selection (figure 3.2.2D) can lead to a reduction in diversity around a selected site due to a genetic hitchhiking effect (Braverman et al. 1995 [7]; Kaplan et al. 1989 [40]; Maynard and Haigh 1974 [56]). The other observable fact in the heat map of correlations is that for different i, the correlations of X_i and X_{n-i} are not the same. However, this does not cause any problem in our model, since the PPD model defines a new variable $Z_i = X_i - X_{n-i}$ whereby any level of correlation between X_i and X_{n-i} can be skillfully cancelled.

3.3.2 *Maximum likelihood estimation of parameters*

MCLE of Zhu and Bustamante's (2005 [95]) Composite Likelihood Model:

Under the assumption of independence among components of the SFS, Zhu and Bustamantes (2005 [95]) CLRT performs very well. How well does it perform when the above assumption is violated? To explore this issue, we simulated 1000 SFS with sample size $n = 10$, $\theta = 30$, $\gamma \in \{-10, -8, -5, -2, 0, 2, 5, 8, 10\}$ and the covariance among components of the SFS, *cov*, in the level of $\{0, 1, 5\}$ by $ppdsfs$ program. We then run CLRT (Zhu and Bustamante 2005 [95]) on each set of the SFS. Figure 3.3A and B plots the MCLE of γ and θ with respect to the level of selection pressure explored. It is clear that when the components of the SFS are independent of each other, the MCLE of both parameters are accurately estimated. However, when the assumption is violated, both estimates deviate from the true value. The magnitude of the deviation is greater when the correlation is higher. To solve this problem, we developed the PPD model in this paper which can cancel

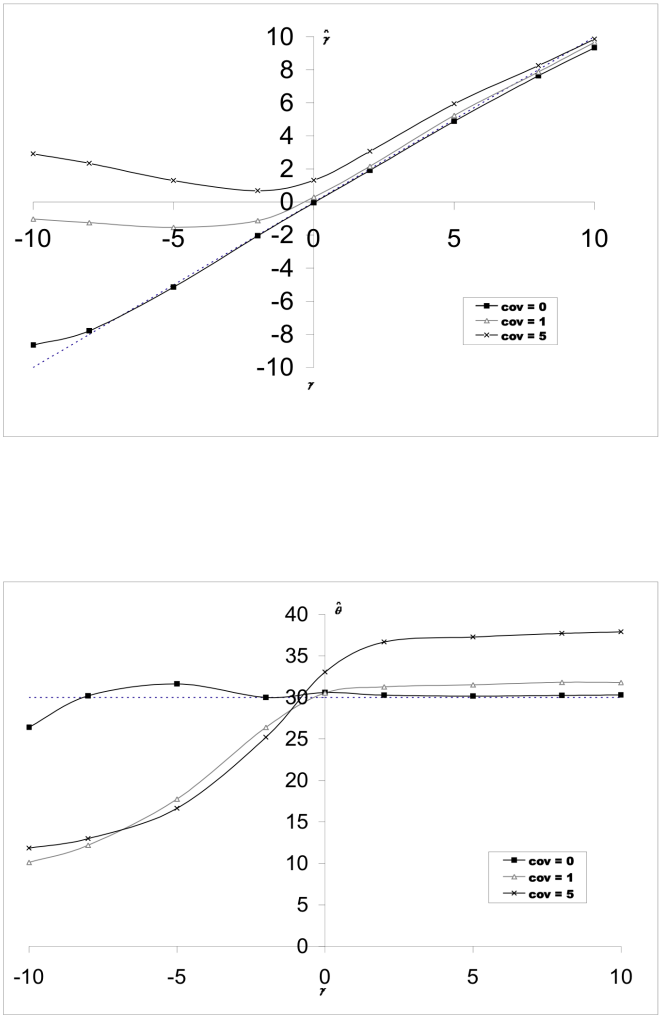


Figure 3.3: *Maximum composite likelihood estimation (Zhu and Bustamante 2005 [95]) of selection coefficient (Top: A) and mutation rate (Bottom: B) for simulated SFS with independent components (cov = 0) and different level of covariance (cov = 1, 5) among pairs of the components. Dash line is the true parameter values that were used for simulation. Data were drawn from ppdsfs program with sample size* $n = 10$, $\theta = 30$, $\gamma \in \{-10, -8, -5, -2, 0, 2, 5, 8, 10\}$.

any strength of the correlation among components of the SFS. The performance of its maximum likelihood estimation of parameters is compared with Zhu and Bustamantes MCLE and evaluated below.

PPD model: The log-likelihood function in the PPD model has two parameters, mutation rate (θ) and selection coefficient (γ). We use Downhill simplex optimization method to find the maximum likelihood estimators for both parameters simultaneously. To increase the probability of finding the global maximum, we started from different initial points which mostly cover the possible ranges of the parameter spaces. The convergence rate (i.e. the percentage of runs which start from different initial points that result in no change in the maximum likelihood estimators) is higher than 99%. To evaluate the performance of the PPDM in estimating mutation and selection parameters, we apply PPDM on each replicate of the SFS, the same simulated data set as in previous MCLE section. Figure 3.4 shows the average $\hat{\gamma}$and $\hat{\theta}$ (MLEs of γ and θ, respectively) from above simulated data sets at different selection coefficients explored. Compared with the CLRT in the PRF model, our PPD model does greatly improve the accuracy in estimating both parameters and the performance of the PPDM is irrespective to the level of the correlation which demonstrate the nice property of our PPD model. Moreover, from Figure 3.4A, we can see that PPDM does an excellent job in finding the $\hat{\gamma}$ when the sequences were sampled from a single population that undergoes either positive selection or neutrality. Similarly, in Figure 3.4B, we can see that PPDM performs reasonably well in estimating θ under positive selection and weak negative selection. However, for strong negative selection, it performs poorly in estimating both γ and θ. Will large sample size improve the performance of the parameter estimation? We simulated 1000 SFS by ppdsfs program with $n = 48$,

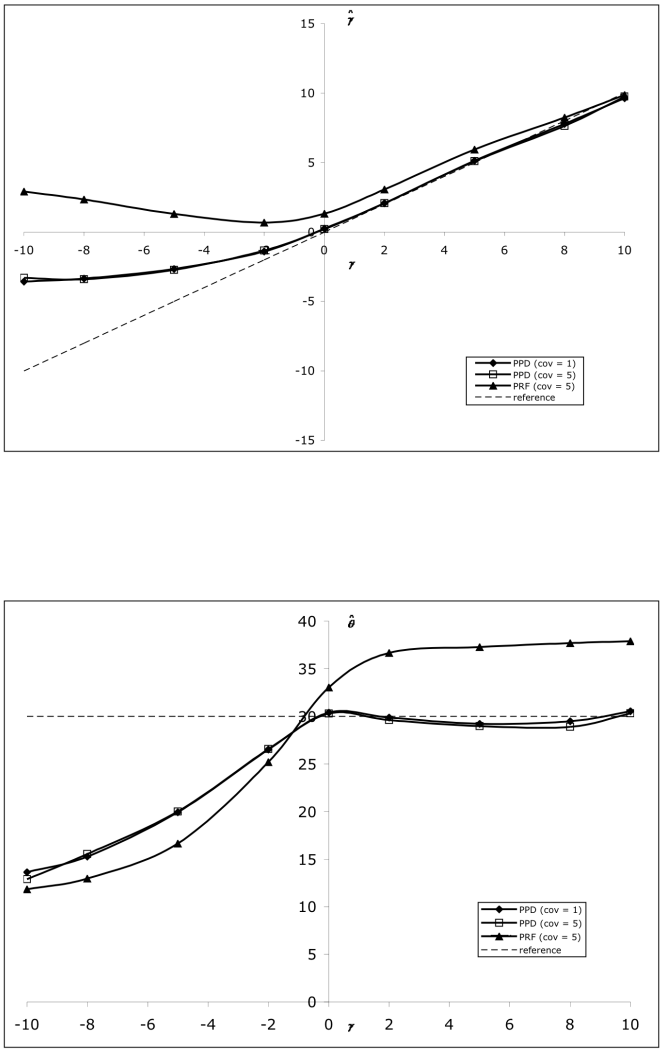


Figure 3.4: *Comparison of Maximum likelihood estimation of parameters between PRF model (Zhu and Bustamante 2005 [95]) and PPD model (Top: A; Bottom: B). Same data set as in figure 3.3.*

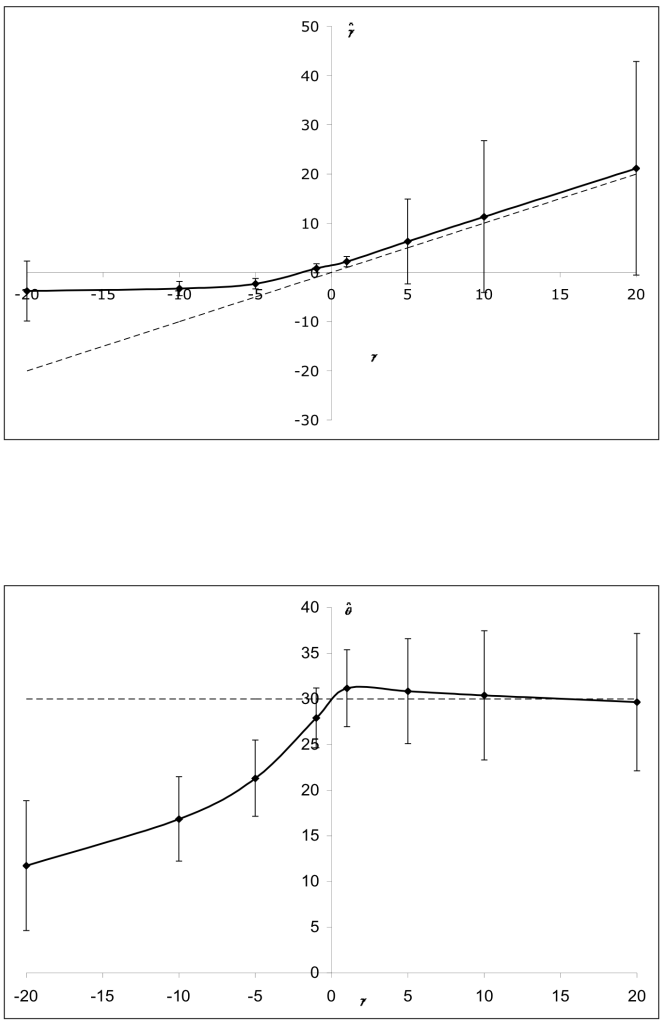


Figure 3.5: *Maximum likelihood estimates of γ (Top: A) and θ (Bottom: B) in the PPD model for simulated SFS with $n = 48$, $\theta = 30$, $\gamma \in \{-20, -10, -5, -1, 0, 1, 5, 10, 20\}$ by ppdsfs program. Dash line is the true parameter value, vertical bars are the standard deviations.*

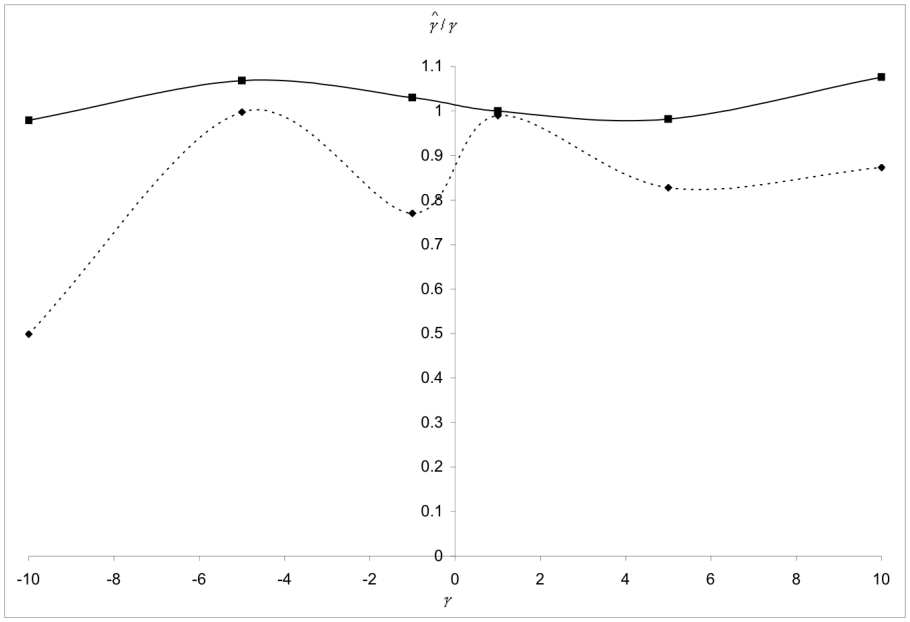


Figure 3.6: *Ratio of the maximum likelihood estimation $\hat{\gamma}$ to the true selection coefficient γ. Solid line is for the estimation given known mutation rate (θ); dashed line is for the profile γ estimation co-estimated with θ.*

Table 3.1: Average SFS(top) and PPDSFS (bottom) under different levels of selection coefficient with $n = 48$, $\theta = 30$.

SFS								
γ	X_1	X_2	X_3	X_4	X_5	...	X_{46}	X_{47}
-20	16.520	4.659	1.717	0.727	0.281	...	0	0
-10	21.735	7.770	3.706	1.997	1.125	...	0	0
-1	29.534	14.172	9.212	6.673	5.226	...	0.212	0.208
0	30.081	15.193	10.036	7.481	6.147	...	0.674	0.608
1	30.238	15.483	10.379	8.062	6.411	...	1.397	1.374
10	30.758	15.707	10.273	8.263	6.687	...	7.859	9.075
20	30.230	15.806	10.570	8.178	6.609	...	10.999	13.941
PPDSFS (Z vector)								
γ	Z_1	Z_2	Z_3	Z_4	Z_5	...	Z_{23}	Z_{24}
-20	16.520	4.659	1.717	0.727	0.281	...	0	0
-10	21.735	7.770	3.706	1.997	1.125	...	0.001	0.001
-1	29.326	13.960	8.984	6.412	4.981	...	0.067	0.726
0	29.473	14.519	9.396	6.792	5.446	...	0.109	1.273
1	28.864	14.086	8.968	6.612	4.989	...	0.031	1.866
10	21.683	7.848	3.333	2.004	1.255	...	0.113	2.481
20	16.289	4.807	1.408	0.661	0.344	...	-0.013	2.579

$\theta = 30$, $\gamma \in \{-20, -10, -5, -1, -, 1, 5, 10, 20\}$ and apply PPDM on each replicate of the SFS. Figure 3.5 plots the mean and standard deviation of $\hat{\gamma}$ and $\hat{\theta}$ at the above range of γ. We can see that larger sample size does not improve the performance of the parameter estimation. Moreover, the variance of the MLEs increases with the selection coefficient for both parameters. The reason for this becomes clear when one examines the effect of selection pressure on the distribution of the site-frequency spectrum, X, and hence the distribution of the Poisson pairwise site-frequency spectrum Z vector. In table 3.1, we present the average of the components of X (top SFS) from 1000 replicates of simulated SFS (by $ppdsfs$ program) with sample size $n = 48$, $\theta = 30$, and $\gamma \in \{-20, -10, -1, 0, 1, 10, 20\}$, and the average of the components of the corresponding PPDSFS (Z vector, bottom). We can see from table 3.1 that for $\gamma < -1$, due to the effect of the negative selection on the SFS which results in $X_i \approx 0$ for $i > n/2$ (in this example, $i > 24$), X_i, which is defined as $X_i - X_{n-i}$, is close to or equal to X_i. Therefore, though we designed the model to take the full vector of the SFS into account and create a new vector Z which cancel the correlation among the components of SFS, if X_{n-i}s are all zero or close to zero, Z vector is no more than just first half vector of the original SFS. In this case, our model loses the power of detecting the signature of the natural selection. Likewise, it also leads to the biased estimation of parameter values. These results suggest that the analysis of single genes using this method may lack of power if genes undergo weak or strong negative selection but has excellent power when genes are subject to positive selection. To improve the performance of the parameter estimation in the negative selected case, we try to calculate the maximum likelihood estimator of selection coefficient conditional on the true mutation rate. MLE of γ is dramatically improved by this way (figure

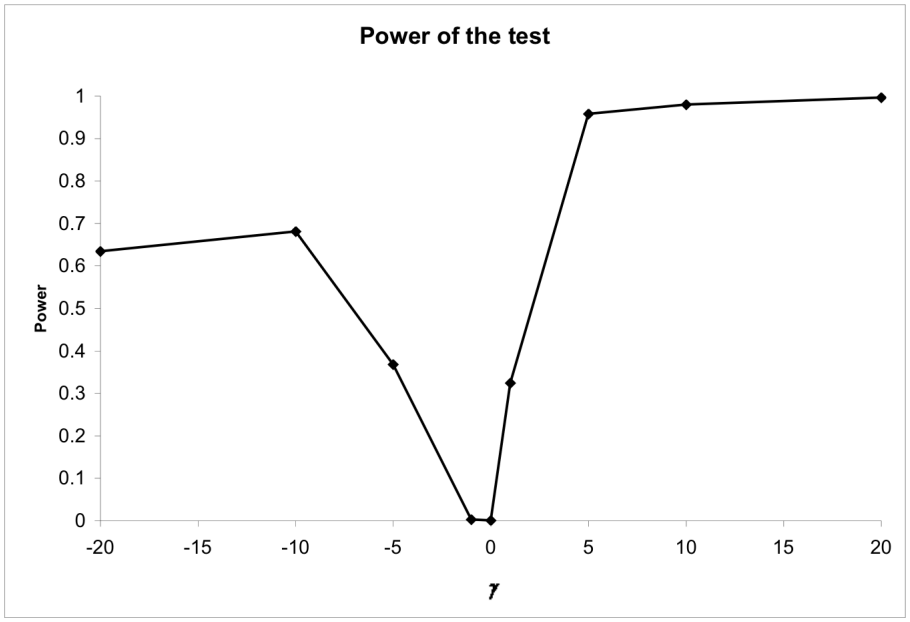


Figure 3.7: *Power of the log-likelihood ratio test in the PPD model. Data were simulated by pddsfs program with* $n = 48$, $\theta = 30$, $\gamma \in \{-20, -10, -5, -1, 0, 1, 5, 10, 20\}$.

3.6). This provides us a good way to modify our maximum likelihood estimation. If one can estimate the mutation rate of a genomic region independently, applying this PPDM may provide us a rather accurate γ estimation.

3.3.3 *Size and Power of the PPDM LRT*

In figure 3.7 we summarize the results for the analysis of the size and power of the PPDM log-likelihood ratio test for simulated data with $n = 48$, $\theta = 30$, $\gamma \in \{-20, -10, -5, -1, 0, 1, 5, 10, 20\}$. We can see from figure 3.7 that the LRT has desired size ($\alpha = 0.05$) for data from neutrality. It has excellent power to detect

positive selection against neutrality and moderate power to detect the signature of negative selection. The reason for this is related to the issue of maximum likelihood estimation of parameters in general in the PPDM framework described above, that is, too few information in Z vectors for data from negative selection results in less accurate parameter estimation, hence reduce the power of the likelihood ratio test.

3.3.4 *Parameter estimation and Power of the PPDM LRT for data from FISHER*

It is important to note that the data for the size and power results presented in Figure 3.7 were drawn from $pddsfs$ program, in which all the assumptions of the PPD model are satisfied. That is, assumptions in the standard PRF model are satisfied with one exception (independence among components of the SFS is relaxed). We are also interested to know how well the PPDM LRT performs on the data from FISHER (Zhu and Bustamante 2005 [95]), which results in different patterns of the PPDSFS due to the different sampling algorithms (figure 3.2). In figure 3.8, we summarize the ratio of the MLEs of mutation and selection parameters to the true parameter values as a function of recombination rate, R. Unfortunately, PPDM LRT is not robust to deviations from the assumption of independent among sites. The performance of the parameter estimation improves with the higher recombination rate. So does the power of the PPDM LRT (Figure 3.9). However, the overall performance of the PPDM LRT on the PPDSFS from FISHER is still not as good as that from the $ppdsfs$ program. Statistical reason for this poor performance may due to the fact that data from FISHER are not so strictly follow all assumptions in the PPD model. While data from $ppdsfs$ program perfectly represent the expected PPDSFS in the population genetic theory (figure 3.2). Bio-

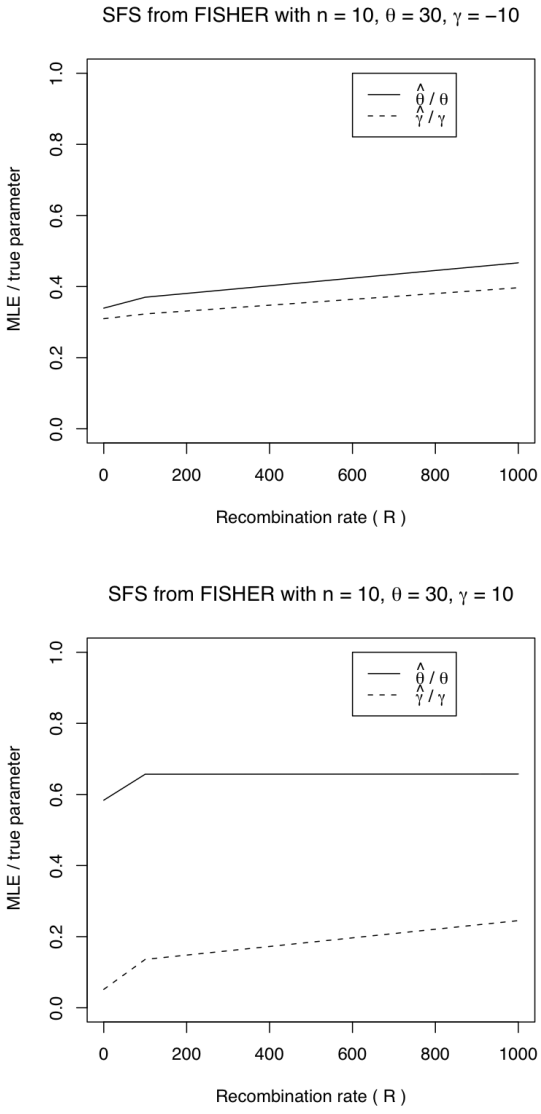


Figure 3.8: *The ratio of MLEs to the true parameter values under negative selection (Top: A. $\gamma = -10$) and positive selection (Bottom: B. $\gamma = 10$), respectively, with different level of recombination rate (R). Data were drawn from Zhu and Bustamantes FISHER program with sample size $n = 10$ and mutation rate $\theta = 30$.*

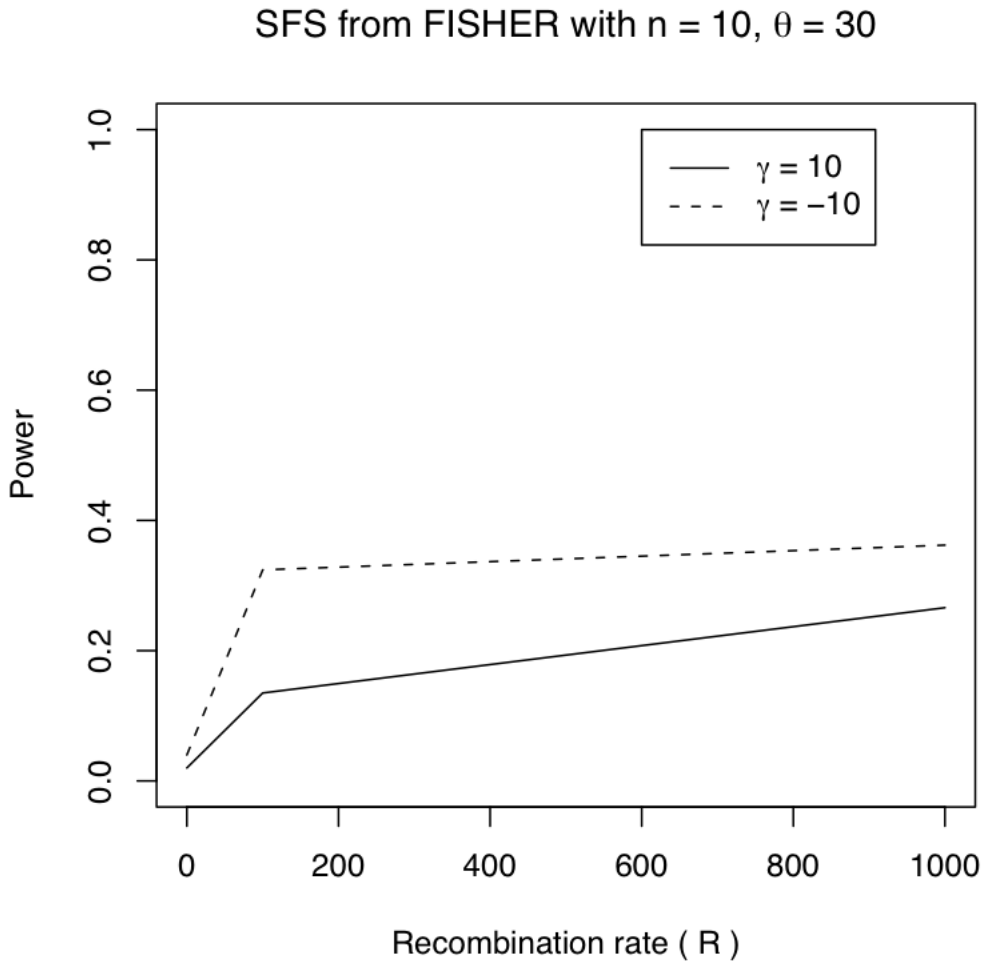


Figure 3.9: *Power of the LRT increases with the recombination rate. Data sets are the same as that in figure 3.7.*

logically, this also indicates that correlation among the components of the SFS that due to the combined effect of linkage and selection can not be well modeled in our PPD model. But it still works reasonably well for detecting positive selection and estimating mutation rate under regular assumption of independent among sites.

3.3.5 *Inference for other population parameters*

One other nice aspect of our PPD model is that one can easily generalize the procedure of PPDM in estimating other population parameters. For example, if one is interested in estimating the population growth rate in a single neutral population, the function $F(i, \gamma)$ can be replaced by a function of the growth rate, $G(i, \beta)$, which could be derived in a proper way based on a certain growth model and the change of allele frequencies. Applying the same procedure of our PPDM on the observed SFS, one could get the maximum likelihood estimation of the growth rate. LRT of constant population size hypothesis can also be performed. Likewise, if one is interested in co-estimating the grow rate and selection coefficient, $F(i, \gamma)$ could be replaced by a function of both of these two parameters, $F(i, \gamma, \beta)$. Using downhill simplex optimization method in 3-dimensional space would provide us the MLEs for all parameters — mutation rate, selection coefficient and growth rate. Similar algorithm could be applied in estimating other population parameters that one is interested.

3.4 *Conclusions*

The Poisson pairwise difference method (PPDM) provides an approach to detect natural selection from arbitrary correlated site frequency spectrum under the model assumptions. It has great power to detect positive selection and performs very well

in estimating both mutation rate and selection coefficient when data are from a single population which undergoes positive selection. SFS is one way of summarizing the polymorphism in sampled sequences. In the PPD model, PPDSFS is the further level of statistic which summarizes SFS. In general, by summarizing data, one may always lose some information compared to the original data set. In our PPD model, if data were from a single negative selected population, PPDSFS is very close or equal to the first half part of the original SFS, therefore, not enough information can be drawn from the PPDSFS with respect to the issue of canceling the pairwise correlation between the components of SFS. Hence, PPDM is not a proper method to detect negative selection. However, for negative selected cases, if one can estimate mutation rate accurately and independently, with given estimated $\hat{\theta}$, our PPDM could greatly improve the accuracy of the γ estimation. The other important issue is that for odd sample size, PPDSFS does not include the key middle component, $Z_{n/2}$, which results in exactly the same PPDSFS for both positive and negative selection with the same magnitude of selection pressure. Therefore even sequences are suggested in the PPD analysis in order to determine the direction of the natural selection.

We also want to point out here that the assumption of the constant panmictic one single population should be satisfied in order to perform the PPD LRT to detect selection. The reason is that some other evolutionary histories besides natural selection, for example, demographic factors, will mimic the effect of the natural selection, hence results in the similar pattern of the SFS and leads to the false rejection of the null hypothesis of neutral evolution. One possible solution of this problem could be explicitly model selection and demography in future in the revised version of the PPD LRT, that is, replace $F(i, \gamma)$ by $F(i, \gamma, \beta)$ in the model.

In particular, partitioning the factors that cause the variation in the SFS into mutational, selective and demographic components, one may get these parameters estimated simultaneously. If recombination is involved, then explicit formula of all parameters is even more complicated but it is worth exploring.

For the point of improving the accuracy of parameter estimations under negative selection and increasing the power of the LRT, we would like to develop, in the future, a proper multivariate distribution of the SFS to model the population evolutionary forces. Such a model would really take the full information in the SFS and the correlation among components into account while does not lose any more information. We will expand upon these methods in subsequent publications.

Chapter 4

A Flexible and Efficient Approach for Estimating Recombination Rate Variation from Population Genomic Data

4.1 *Introduction*

Efficient methods for association mapping presuppose that one has accurate information on the local rate of recombination along the genome. Classical methods for estimating recombination rates from natural populations include pedigree studies, sperm typing analysis and methods based on predictions from population genetics. In humans, the difficulty of obtaining large pedigrees limits the utility of pedigrees to estimation of large-scale recombination rates (Kong et al. 2002 [49]). Likewise, while sperm typing can provide accurate estimates of the local recombination rate in male gamete production, it is very labor intensive and expensive. These limitations coupled with the increasing availability of genome-wide polymorphism data from humans and other species make estimation of recombination rates via population genetic theory an attractive alternative. A number of population genetic estimators of the population recombination rate ($R = 4N_e r$, where r is the rate of crossing over for the region and N_e is the effective population size) are currently available, including moment-based estimators (Hudson, 1985 [28]; Hey and Wakeley 1997 [25]; Wall 2000 [85]), full maximum likelihood estimators (Griffiths and

Marjoram 1996 [23]; Kuhner et al. 2000 [52]; Nielsen 2000 [64]; Fearnhead and Donnelly 2001 [13]) and approximate likelihood estimators (McVean et al. 2004 [59]; Hudson 2001 [32]; Fearnhead and Donnelly 2002 [14]; McVean et al. 2002 [58]; Li and Stephens 2003 [55]; Crawford et al. 2004 [9]; Fearnhead et al. 2004 [15]; Fearnhead and Smith 2005 [16]). Since the effective population size is confounded within the estimate of the recombination rate, population genetic estimators are by definition dependent on assumptions regarding the demographic history of the sample. A limitation of many of these approaches, therefore, is that they are based on the assumption that the population under study is randomly mating and constant in size an assumption violated by nearly all populations to which the approach is applied.

In this paper, we present a novel statistical method for estimating the population recombination rate via multiple linear regression (MLR) and non-parametric bootstrap. Three advantage of our method are that (1) it can readily accommodate complex demographic history, (2) provide confidence intervals for the estimated recombination rate, and (3) is computationally efficient and applicable to whole-genome data. Furthermore, since the method weighs heavily the variance of new mutations in estimating recombination rates, it may be able to detect recent changes in recombination rate that do not leave an explicit LD signal.

Our method is based on a readily discernible statistic of the data: the observed variability in the number of mutations at different frequencies across sub-samples of the data. It is important to note that the idea of using the variance of mutation counts in a sample to estimate recombination rates is not new. Nearly two decades ago, Hudson (1987 [30]) introduced an estimator of the population recombination rate based on the sample distribution of pairwise differences. Since then, much

has been done to improve upon Hudsons pioneering work. For example, Wakeley (1997 [81]) proposed an improved version of Hudsons (1987 [30]) estimator that has smaller bias and standard error. As we show in the supplementary information, Hudsons estimator is closely related to the regression estimator discussed below.

Formally, the site-frequency spectrum (SFS) summarizes single nucleotide polymorphism frequency data for a genomic region in terms of a vector X such that X_i is the number of SNPs at frequency i out of n in the sample where n is the number of chromosomes sequenced. For example, X_1 is the observed number of singletons in the sample or the number of SNPs at frequency 1 out of n. For the standard neutral Wright-Fisher model of population genetics, Fu (Fu 1995 [21]) derived the expected value, variance and covariance of each component of the SFS in both folded (the ancestral state for each single nucleotide polymorphism is unknown) and unfolded (assuming the ancestral state is known) cases assuming complete linkage among sites. Across independent realizations of the evolutionary process, X will vary stochastically so that for each component one has an associated variance V_i. For example, V_1 is the variance in the number of singletons that one would observe if one were to have sampled a different set of individuals. Here we describe how recombination affects the variances and co-variances of the components of the SFS (SFS variances) in a fully predictable way and how by estimating SFS variances, one can predict the recombination rate of a genomic region for a given demographic model. A major advantage of this approach is that it does not require calculation of pair-wise linkage disequilibrium and, thus, does not require phasing of the data.

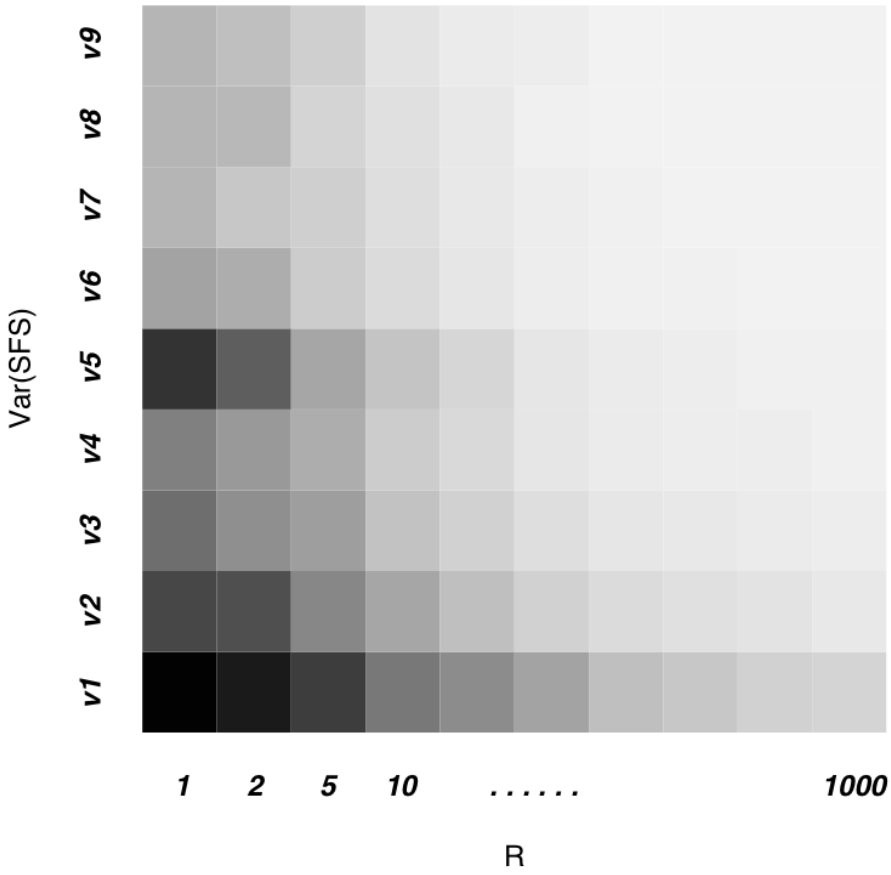


Figure 4.1: *Variances of the SFS components decrease when recombination rate increases. For* 10000 *replicates of simulated data sets, each with* $n = 10$, $\theta = 30$, *and* $R \in \{1, 2, 5, 10, 20, 50, 100, 200, 400, 1000\}$.

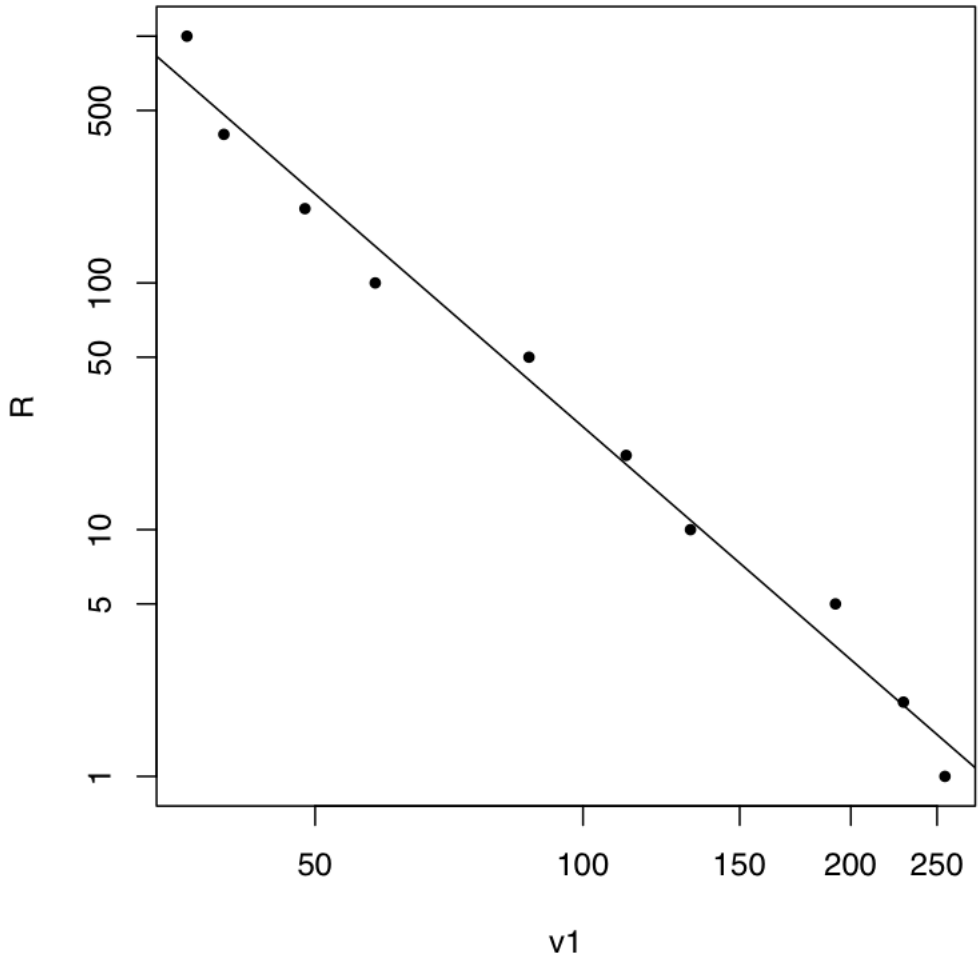


Figure 4.2: *Linear regression of log transformed recombination rate (logR) and log transformed variance in the number of singletons in the sample, log(V_1) under the standard neutral Wright-Fisher model. Each point represents the average of* 1000 *data points.*

Table 4.1: Correlation matrix of variances of SFS components for $n = 10$, $\theta = 30$, $R \in \{1, 2, 5, 10, 20, 50, 100, 200, 400, 1000\}$

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9
V_1	1.000	0.9886	0.9879	0.9881	0.9564	0.9726	0.9829	0.9711	0.9803
V_2	0.9886	1.000	0.9850	0.9912	0.9794	0.9923	0.9815	0.9943	0.9905
V_3	0.9879	0.9850	1.0000	0.9984	0.9792	0.9850	0.9994	0.9768	0.9928
V_4	0.9881	0.9912	0.9984	1.0000	0.9840	0.9918	0.9977	0.9864	0.9974
V_5	0.9564	0.9794	0.9792	0.9840	1.0000	0.9955	0.9789	0.9880	0.9876
V_6	0.9726	0.9923	0.9850	0.9918	0.9955	1.0000	0.9838	0.9976	0.9957
V_7	0.9829	0.9815	0.9994	0.9977	0.9789	0.9838	1.0000	0.9747	0.9924
V_8	0.9711	0.9943	0.9768	0.9864	0.9880	0.9976	0.9747	1.0000	0.9923
V_9	0.9803	0.9905	0.9928	0.9974	0.9876	0.9957	0.9924	0.9923	1.0000

4.2 *Results and Discussion*

We first consider the problem of predicting the population recombination rate from polymorphism data arising under a known demographic model. Using standard coalescent algorithms, we simulated 10,000 replicate samples for each of 10 levels of recombination rate $R \in \{1, 2, 5, 10, 20, 50, 100, 200, 400, 1000\}$ under a fixed mutation rate $\theta = 4N_e\mu = 30$ where μ is the regional mutation rate per chromosome. (These parameter values correspond roughly to a 30 Kb region in humans with recombination rate varying from 2.5×10^{-4} cM to 0.25 cM.) Figure 4.1 shows how variances of SFS components change with recombination rate in this example. The darkness of the square indicates the magnitude of the variance. It is clear that V_i decreases when the recombination rate increases. This pattern is true for

Table 4.2: JMP output for multiple linear regression of $log(R)$ on $log(V_1)$ for 10000 replicates of simulated data set, each with $n = 10$, $\theta = 30$, $R \in \{1, 2, 5, 10, 20, 50, 100, 200, 400, 1000\}$

$Log(R) = 17.695242 - 3.1342004 Log(V_1)$	
RSquare	0.985161
RSquare Adj	0.983306

Analysis of Variance

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Ratio**	**Prob > F**
Model	1	47.080044	47.0800	531.1096	< .0001
Error	8	0.709158	0.0886		
C.Total	9	47.789201			

Parameter Estimates

Term	**Estimate**	**Std Error**	**t Ratio**	**Prob > |t|**
Intercept	17.695242	0.626047	28.27	<.0001
$Log(V_1)$	-3.1342	0.135999	-23.05	<.0001

each SFS component. To model the relationship between V_i and R, should we include all V_is in the model? If V_is are auto-correlated, then subset of V_is should be sufficient. Table 4.1 shows the correlation matrix of V_is for the above example. We found that the variances of the SFS components are highly correlated. That is, as recombination rate increases, the variance of the number of singletons (V_1) decreases in a functionally similar fashion as does the variance of doubletons (V_2), triplets (V_3), and so on. When we perform the multiple linear regression of R on all V_is including all pairwise covariances among SFS components and use both stepwise selection and best subset methods (*Minitab/JMP* software is used for fitting models), all other terms are dropped except V_1 in the model. Scatter plot of R to the average V_1 across simulated data sets shows a curvilinear relationship suggesting that linear regression of log-transformed data could be used to estimate R from a linear combination of the components in V. Using a step-wise addition rule, we find that $log(V_1)$ alone is a sufficient predicator variable for the population recombination rate with the best fit linear regression explaining nearly 98.5% of the variance ($R^2 - adjusted = 0.983$) as shown in figure 4.2. The output of the regression is shown in table 4.2. Diagnostic tests (linearity, constant variance, normality, independence) for validation of the model are checked and all assumptions for performing LR are satisfied (Note: for all regressions performed, diagnostic tests were checked and satisfied, results not shown). This simple example shows that for a fixed level of the mutation rate, the transformed recombination rate and the first component of SFS variances are highly correlated so that given an observed variance of singletons across samples one can accurately predict the recombination rate under which the data were generated.

The nice property of high correlation among variances of the SFS components

also provides great flexibility for estimation since when data on singletons is not available (e.g., due to ascertainment bias) one can use variances of other SFS components to predict R.

Next we explored how to estimate recombination rate variation while controlling for the uncertainty regarding the underlying mutation rate for the study region. We applied the same analysis above on 10,000 simulated data sets with a fixed number of segregating sites $S = 10$, and R in the same range described above. As before, the best-fit linear regression of the transformed recombination rate R as a function of V_1 under the standard neutral model has near-perfect explanatory power ($R^2 = 0.9833$, $R^2 - adjusted = 0.9809$). The linear relationship between $log(R)$ and $log(V_1)$ holds for any given fixed number of segregating sites that we explored in this study ($S = \{10, 20, 30, 50, 100\}$) suggesting great flexible in controlling the resolution of the recombination rate estimation. By choosing a fixed number of segregating sites in a genomic region, without introducing any additional predictor variables in the model, one could easily predict the recombination rate for that region using the observed SFS variances across samples.

For n sampled sequences of real data, however, we only have one observed SFS vector. To estimate the SFS variances, one therefore needs to couple a re-sampling step such as non-parametric bootstrapping to the MLR. We compared the evolutionary and re-sampling variances by simulation and found that non-parametric bootstrap estimates of variances are systematically smaller. However, we can correct this bias by using bootstrapped estimates in the regression model-fitting step as well as in recombination rate prediction for real data step. Once again a high correlation $R^2 = 0.983$ ($R^2 - adjusted = 0.982$) is obtained in the best-fit MLR model for 1,000 single data sets of sample size $n = 60$ with R in the range

as before and fixed number of segregating sites. The SFS variances are estimated by a non-parametric bootstrap strategy by re-sampling 60 sequences with replacement for each data set 5,000 times. To reduce the burden of the computation and without loss generality, we used sub-samples of the bootstrap $n_{sub} = 10$ to fit the model. For this scheme, an $R^2 = 0.9642$ ($R^2 - adjusted = 0.9625$) is obtained for the regression of $log(R)$ on $log(V_1)$ under the standard neutral model. This indicates that even when the covariates need to be estimated from the data, about 96.4% of the variation in the transformed recombination rate can be explained by linear relationship between $log(R)$ and $log(V_1)$. In other words, linear combination of log-transformed SFS variances can be used to reliably estimate the recombination rate in the genomic region of interest. For all subsequent tests, we use the MLR with re-sampling as one would with a real application of the method.

To evaluate the performance of the MLR in estimating recombination rate variation across a region, we applied the method on 1000 simulated data sets with a hotspot located between scaled region $0.4 \sim 0.5$ in which recombination occurs at a rate 10 times greater than the background (Li and Stephens 2003 [55]). In the procedure of predicting the local recombination rate, which is described in the method section (D), we fix the window size of 10 SNPs and slide windows along the sequences by 1 SNP each time. For each region with length 0.1 starting from scaled position 0 to 1, we estimate the average predicted recombination rate. Figure 4.3 shows the 95% range and mean ratio of the estimated recombination rate to the true background parameter value along the sequences for replicate data sets with (dark grey) and without (light grey) a hotspot. In Figure 4.3a, where the data is simulated under the standard neutral model and standard neutral model is assumed in the MLR procedure, the ratios are around 1 for data with uniform

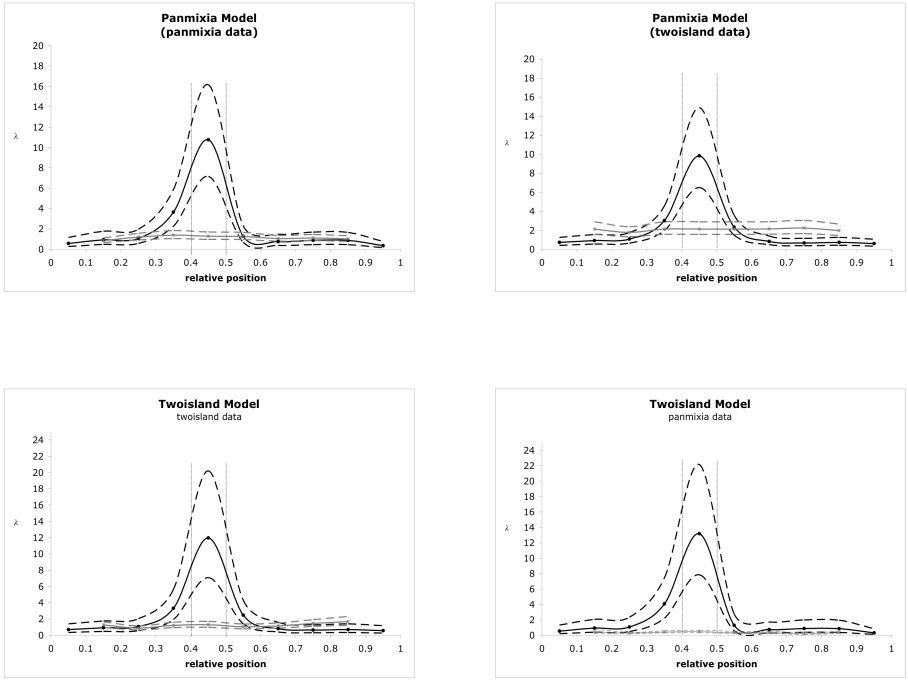


Figure 4.3: *The ratio (λ) of recombination rate estimates to the background parameter values. Data were simulated via ms and post-processed using Li and Stephenss program (2003 [55]). Top Left (a). Data from a single population of constant size (panmixia) and panmixia assumed when estimating recombination; Top Right (b). Data simulated under two-island population model, but fit assuming panmictic population; Bottom Left (c). Data simulated under and fit assuming a two-island population structure model; Bottom Right (d) Data from standard neutral model, but fit assuming two-island population structure model. Dash lines are 95% confidence upper and lower bounds. Solid lines are estimated means. Dark lines corresponding to data with hotspot on the known region,* $0.4 \sim 0.5$ *(region between vertical dash bars), with magnitude 10 times greater than background recombination rate; gray lines are for data with uniform recombination rate along the whole region. Window size* $w = 10$ *SNPs.*

recombination rate, around 10 for hotspot data as expected. We see that the approach performs very well when the correct demographic model is assumed in that it correctly identifies the hotspot position as well as the magnitude. Compared with the method in Li and Stephens (2003 [55]), our method does not require any prior distribution of the magnitude or density of recombination hotspots.

A potential problem for all methods that aim to estimate the population recombination rate is misspecification of the demographic model. In order to investigate this issue, we applied our approach on data simulated under a two-island model with moderately high migration ($4N_em = 4$,$n = 30$ sequences sampled from each sub-population, consistent with the data used in Li and Stephens (2003 [55])). The results are mixed. On the one hand the location of the hotspot is accurately ascertained. However, misspecification of the demographic model leads to a slight over estimation of the background recombination rate in the case we studied (light grey in figure 4.3b). However, in comparing the estimates of the recombination rate within the hotspot region and that in the flanking regions we find that the method is still able to detect a 10-fold increase in recombination (dark grey in figure 4.3b).

One advantage of our approach is that it is readily amenable to incorporation of complex demography by simply simulating under the desired model during the MLR step. When applying the approach to data simulated under the two-island model discussed above, the regression continues to have very high explanatory power ($R^2 = 96.39\%$, $R^2 - adjusted = 95.48\%$). Figure 4.3c shows the results from the data in two-island model with known hotspots between $0.4 \sim 0.5$ (dark grey) and without hotspot (light grey). It is clear that after taking population structure into account, the method of predicting recombination rates from the variances

of the SFS components performs very well. Likewise, figure 4.3d demonstrates that analyzing data from standard neutral population by two-island MLR model results in downward biased estimation of the parameter (light grey in figure 4.3d), however, correcting the biased background leads to accurate estimation of the ratio and identification of the hotspot location (dark grey in figure 4.3d).

We have explored the effect of several other demographic scenarios on estimating recombination rate and localizing hotspots. For example, we have investigated the performance of our method under one-island, exponential growth, severe and weak bottleneck scenarios with and without recombination hotspots. For each model, we simulated 1000 replicate data sets, with 60 sequences per replicate and fixed number of segregating sites $S = 50$. A hotspot is located in the same region $(0.4 \sim 0.5)$ with magnitude 10 times greater than the background rate as before. Data from one-island population assumes that all 60 sequences were sampled from one single sub-population, with migration rate $Nm = 4$ to another sub-population. Samples from exponential growing population assume growth rate $G = 3.2$. For recent bottleneck scenario, we assume the bottleneck population size was in fraction f of its current size between $0.25N_e$ and $0.5N_e$ generations ago, and prior to that, the population size was the same as the current population size, $f = 0.1$ for sever bottleneck and $f = 0.9$ for weak bottleneck. Figure 4.4 shows the mean ratio of estimated recombination rate to the background rate along the region for these four demographic scenarios. We see that misspecification of the underlying demographic model can lead to biased estimation of the recombination rate in these cases, with parameter values downwardly biased if we incorrectly assumed that the sampled data arise from a standard neutral model with constant population size. As before, the estimated location of hotspot appears to be robust to the underlying

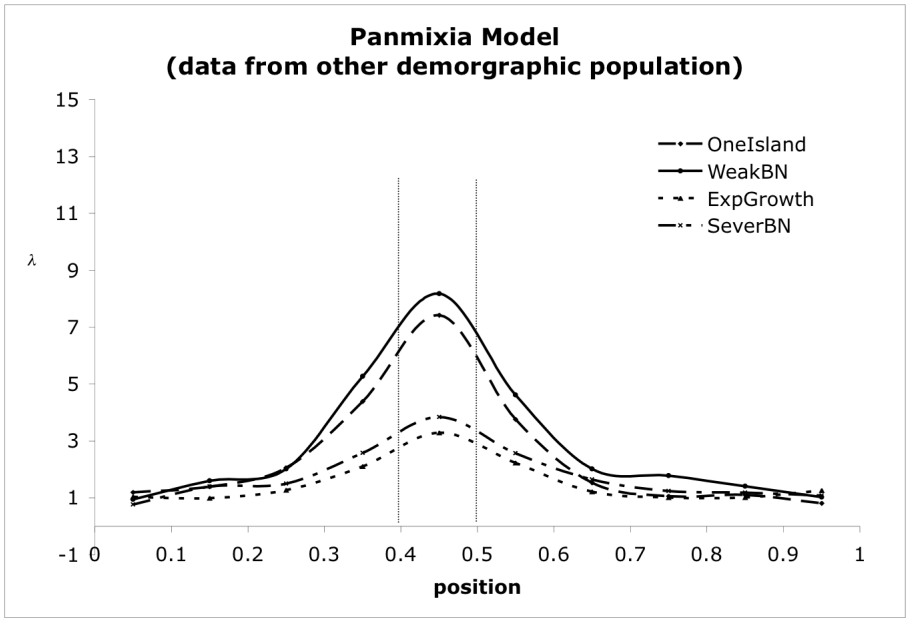


Figure 4.4: *The ratio (λ) of recombination rate estimates to the background parameter values under various demographic models. Data simulated under one-island, exponentially growing, weak and severe bottleneck scenarios ms and post-processed using Li and Stephenss program (2003 [55]). Data were analyzed by multiple linear regression that is fit assuming a standard neutral model with constant population size.*

demographic model.

It is important to note that the choice of the window size on the regression region may affect rate estimation. Windows are mostly overlapped when we move one SNP site step by step. If the window size is too large, rate estimates are upward or downward interfered by adjacent SNPs, especially when the window ranges from no or low recombination rate region to a hot spots region. From our study, the suggested window size is between $10 \sim 20$ SNPs.

We have also used our approach to estimate fine-scale recombination rate variation near the well-characterized $TAP2$ recombination hotspots in the human genome (haplotype sequences were kindly provided by Sir Professor Alec J. Jeffreys). A total of 60 sequences with 48 SNPs were included in the analysis. The recombination rate between adjacent pairs of SNPs (as well as associated prediction intervals) was estimated using a sliding window approach (10 SNPs in each window). Figure 4.5 shows the lower bound of the 95% prediction interval of the recombination rate along the $TAP2$ genomic region. The hot spots regions identified by our approach are completely consistent with the results from both sperm typing and haplotype analysis (Jeffreys et al. 2000 [36]). That is we detect a strong signals of dramatically active recombinational exchange in the regions between markers $T15(4017)$ and $T18(4553)$, $T23(4917)$ and $T24(4934)$, and $T27(5188)$ and $T30(5417)$.

Lastly, we have also applied this approach on a 206-kb region on human chromosome $1q42.3$, which contains a well-characterized autosomal crossover hotspots around the highly variable minisatellite MS32 (Jeffreys et al. 1998 [35]). For this analysis, 80 individuals with 214 SNPs were included (10 SNPs in each window). Figure 4.6a shows the mean ratio of predicted recombination rate to the estimated

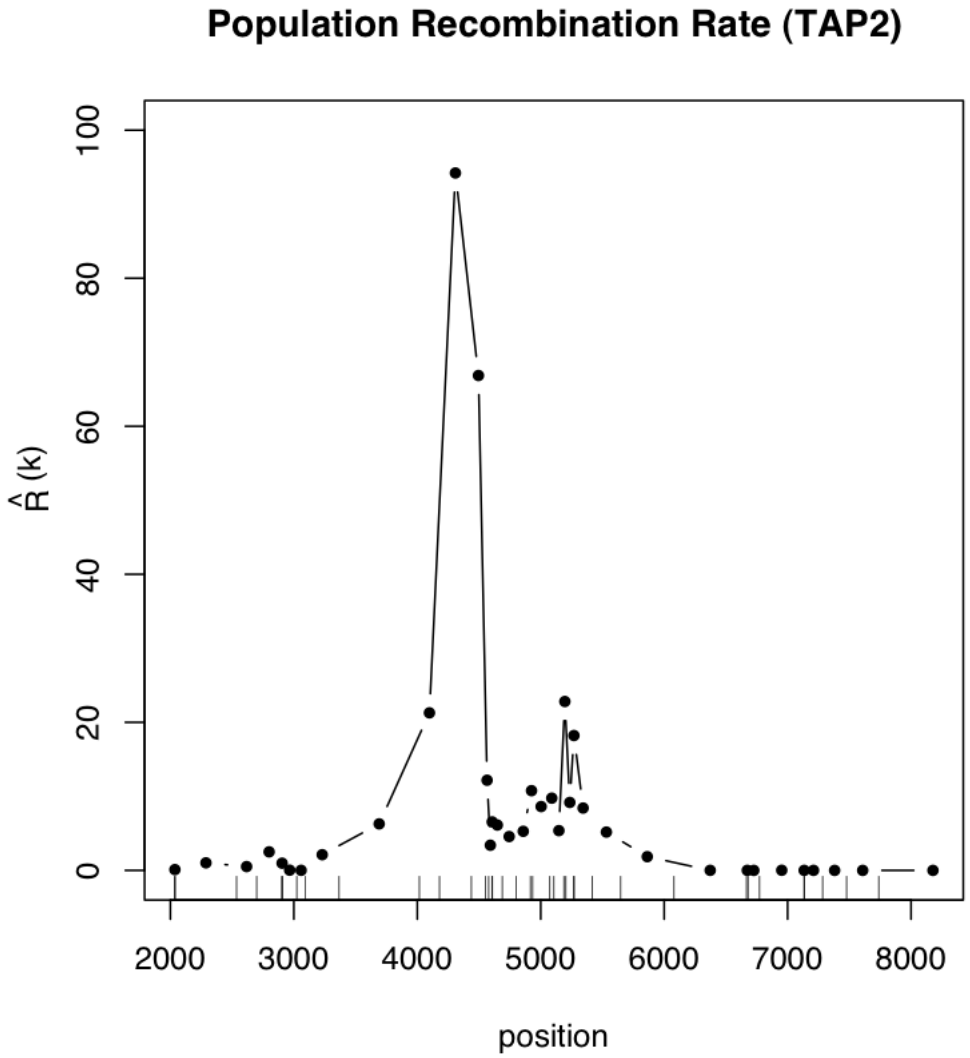


Figure 4.5: *The lower bound of 95% prediction interval of recombination rate along the TAP2 region. SNPs marker positions are consistent with those in Jeffreys et al. (2000 [36]).*

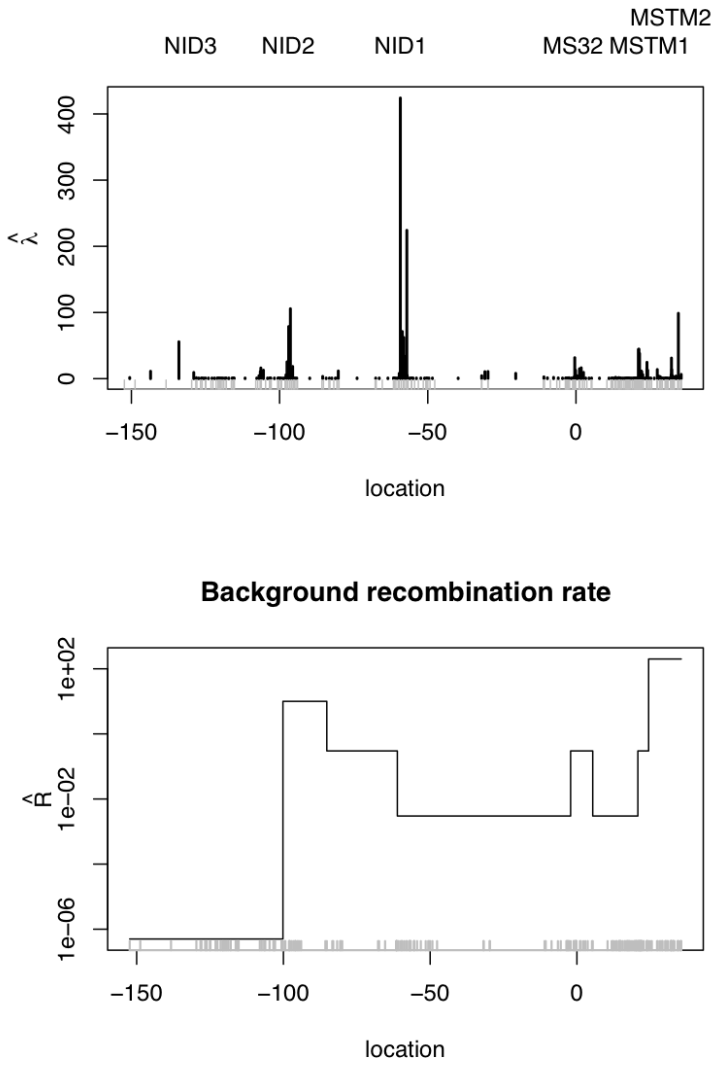


Figure 4.6: *Top (a). Ratio of recombination rate estimates to the background values in the 206 kb interval surrounding minisatellite MS32 on chromosome* 1q42.3. *Bottom (b). Estimated background rate along the region.*

background rate (the estimated background rates along the region is shown in figure 4.6b). We successfully identified the putative hotspots termed *NID1*, *NID2*, *NID3* in and near the *NID* gene, as well as *MS32*, *MSTM1* and *MSTM2* which are consistent with that from sperm crossover analysis and Fearnheads method (Fearnhead et al. 2004 [15]). It is interesting to note that our approach is able to identify the *NID3* and *MS32* hotspots, which the LDhot method based on composite-likelihood of pairwise sites does not seem to find (McVean e al. 2004 [59]; Jeffreys et al. 2005 [37]). The simplest explanation for this is that active hot spots such as *NID3* and *MS32* have evolved so recently such that they are too young to leave full mark on haplotype diversity in the population (Jeffreys et al. 2005 [37]). It has been suggested that LD based approaches for estimating recombination rate may be unable to find very recent hotspots since recombination LD decays slowly. One reason that our approach may be able to work on detecting such hotspots is that we rely heavily on information from singletons in the sample, which are often the most recent mutations in the data.

Applications above assume that all SNPs are evolving neutrally. However, selection does affect the recombination rate prediction, since it has similar effect on the variances of SFS components as recombination does. For example, a recent selective sweep wipes the variation and mimics the effect of recombination. Because sweep also reduces the number of segregating sites (S) and pairwise differences (π), while recombination on average should not affect these two statistics, one possible way to distinguish these two factors is that one can simulate neutral data with estimated recombination rate and compare S and π with those from observed data respectively. If the differences are significant, sweep may be the main cause of variation reduction.

4.3 *Conclusion*

While the algorithm we have presented is fast, flexible, and scalable to the whole genome level, a few caveats must be raised. In order to make inference, we must still presuppose some demographic model for the data. Our preliminary results confirm the predictions of population genetic theory in that recombination rate estimates will be sensitive to the demographic model used in the MLR fitting step. This sensitivity is not likely unique to our approach and probably holds for the majority of algorithms currently in use. At the same time, it also appears that our approach is robust to demography for the problem of detecting recombination rate variation. Secondly, our method does not currently deal with ascertainment bias in the data. Many types of ascertainment biases (such as using a small panel to discover SNPs and then a large panel to genotype) can easily be incorporated into the data simulation step used by the MLR procedure. When ascertainment differs dramatically among SNPs in the same region, however, this may likely cause problems for any method aiming to discover variation in recombination rate.

4.4 *Methods*

Let θ be twice the expected number of mutations at a locus in a population per generation and let X_i be the number of sites at frequency i out of n randomly sampled sequences. For example, if we assume that samples are taken from a population that evolves according to the Wright-Fisher model, that all mutations are selectively neutral and fully linked, then by Fu (1995 [21]), the variance of X_i,$V_i = Var(X_i) = \frac{1}{i}\theta + \sigma_{ii}\theta^2$, where σ_{ii} is a function of i and sample size n.

Two types of the data sets are simulated to explore how variances of X_is change

with the recombination rate ($R = 4N_e r$).

A. **Direct Simulation:**

$Q = 10000$ replicates of data sets are simulated by Hudsons (2002 [33]) *ms* program . Each replicate has $n = 10$ random sequences with given mutation rate, θ, and population recombination rate, R.

B. **Non-Parametric Bootstrapping:**

In reality, for a set of sampled sequences, the variances of the components of the SFS are unknown. However, we can estimate them by bootstrapping. We accomplish this by:

(1) Sample $n = 60$ sequences by Hudsons (2002 [33]) *ms* program with given fixed number of segregating sites $S \in \{10, 20, 30, 50, 100\}$ and $R \in \{1, 2, 5, 10, 20, 50, 100, 200, 400, 1000, 2000\}$.

(2) From above n sequences, re-sample $n_{sub} = 10$ sequences with replacement. Repeat this re-sampling scheme to generate $Q = 5000$ replicates of bootstrap samples.

C. **Multiple Linear Regression Model:**

(1) For each replicate of the data set (n or n_{sub} sequences), calculate the SFS, X ;

(2) Obtain the variance of X_i, V_i, among Q replicates of the data set;

(3) Transform R to $log(R)$ and V_i from step (2) to $log(V_i)$;

(4) Fit a best-fit multiple linear regression model with $log(R)$ as the response and $log(V_i)$ as the possible set of predictors by both stepwise

selection and best subset methods (*Minitab/JMP* software is used for fitting models);

(5) Check all assumptions for fitting a linear regression model, including normality, equal variance of residues and independence assumptions.

D. **Evaluating the performance of the model in predicting the recombination rate:**

Datasets were first simulated under *ms* program (Hudson 2002 [33]) assuming constant recombination rate along sequences, then post-processed according to the algorithm described in Li and Stephens (2003 [55]) to produce a recombination hotspot. Same data files (available from $http : //www.biostat.umn.edu/ \sim nali/software/data/hotspot60.tar.bz2$) have been used for the analysis. It contains 1000 simulated replicates. Each replicate has 60 sequences with background recombination rate $R = 20$ and hotspot 10 times greater. We first perform the non-parametric boostrapping as described in B. On sequences newly sampled from *ms* program (Hudson 2002 [33]) with $n = 60$, find a best-fit multiple linear regression model following the procedure described above in C. For each simulated replicate with known hotspot, we bootstrap the same sample size as in the model fitting step above with a given window size w, which is equal to the fixed number of segregating sites in the MLR fitting. Then,

(1) For each bootstrapping replicate, calculate the SFS, X;

(2) Obtain the variance of X_i, V_i, among Q bootstrapping replicates, get predictor variables by making corresponding transformations for each variable;

(3) Plug predictor variables from (2) into the fitted multiple linear regression model to calculate the prediction intervals (PI) or the confidence intervals (CI) for transformed recombination rate (R).

(4) Slide windows by the given number of skipped SNPs, k ($k = 1$ in this paper), and repeat the same procedure for each simulated replicate sequences with known hotspots.

Appendix

Let $X = [X_1, X_2, X_3, ..., X_{n-1}]$ be the site frequency spectrum (SFS) for sampled DNA sequences with size n, and $Z_i = X_i - X_{n-i}$, show that $E[Z_i|\gamma] = E[Z_i| - \gamma]$, for $i \neq \frac{n}{2}$.

Define $H(i,\gamma) = F(i,\gamma) - F(n-i,\gamma)$ where $F(i,\gamma)$ is defined as equation 1.11. If we define $H(i,\gamma) = F(i,\gamma) - F(n-i,\gamma)$, then

$$H(i,\gamma) = \begin{pmatrix} n \\ i \end{pmatrix} \int_0^1 \frac{1-e^{-2\gamma(1-x)}}{1-e^{-2\gamma}} [x^{i-1}(1-x)^{n-i-1} - x^{n-i-1}(1-x)^{i-1}]dx$$

Therefore,

$$\begin{aligned} E[Z_i|\gamma] &= E[X_i - X_{n-i}] \\ &= E[X_i] - E[X_{n-i}] \\ &= \theta F(i,\gamma) + \theta^2\sigma_{i,n-i} - \theta F(n-i,\gamma) - \theta^2\sigma_{i,n-i} \\ &= \theta[F(i,\gamma) - F(n-i,\gamma)] \\ &= \theta H(i,\gamma) \end{aligned}$$

So we have

$$\begin{aligned} E[Z_i|\gamma] - E[Z_i| - \gamma] &= \theta[H(i,\gamma) - H(i,-\gamma)] \\ &= \theta \begin{pmatrix} n \\ i \end{pmatrix} \int_0^1 [\frac{1-e^{-2\gamma(1-x)}}{1-e^{-2\gamma}} - \frac{1-e^{2\gamma(1-x)}}{1-e^{2\gamma}}] \times \\ &\quad [x^{i-1}(1-x)^{n-i-1} - x^{n-i-1}(1-x)^{i-1}]dx \end{aligned}$$

Define

$$g(x,\gamma) = \frac{1-e^{-2\gamma(1-x)}}{1-e^{-2\gamma}} - \frac{1-e2\gamma(1-x)}{1-e^{2\gamma}}$$

$$f(x,i,n) = x^{i-1}(1-x)^{n-i-1} - x^{n-i-1}(1-x)^{i-1}$$

Then,

$$E[Z_i|\gamma] - E[Z_i| - \gamma] = \theta \begin{pmatrix} n \\ i \end{pmatrix} \int_0^1 g(x,\gamma) f(x,i,n) dx$$

1. Show $g(x,\gamma) = g(1-x,\gamma)$ for $0 < x < 1$

$$\begin{aligned} g(x,\gamma) - g(1-x,\gamma) &= \frac{1-e^{-2\gamma(1-x)}}{1-e^{-2\gamma}} - \frac{1-e^{2\gamma(1-x)}}{1-e^{2\gamma}} - \frac{1-e^{-2\gamma(x)}}{1-e^{-2\gamma}} \\ &- \frac{1-e^{2\gamma(x)}}{1-e^{2\gamma}} \\ &= \frac{e^{-2\gamma x} - e^{2\gamma(1-x)} - e^{-2\gamma(1-x)} + e^{2\gamma x}}{(1-e^{-2\gamma})(1-e^{2\gamma})} \\ &+ \frac{-e^{2\gamma x} + e^{-2\gamma(1-x)} + e^{2\gamma(1-x)} - e^{-2\gamma x}}{(1-e^{-2\gamma})(1-e^{2\gamma})} \\ &= 0 \end{aligned}$$

2. Show $f(x,i,n) = f(1-x,i,n)$ for $0 < x < 1$

$$\begin{aligned} f(x,i,n) + f(1-x,i,n) &= x^{i-1}(1-x)^{n-i-1} - x^{n-i-1}(1-x)^{i-1} \\ &+ (1-x)^{i-1}x^{n-i-1} - (1-x)^{n-i-1}x^{i-1} \\ &= 0 \end{aligned}$$

Lastly,

$$\begin{aligned}
E[Z_i|\gamma] - E[Z_i| - \gamma] &= \theta \begin{pmatrix} n \\ i \end{pmatrix} \int_0^1 g(x,\gamma) f(x,i,n) dx \\
&= \theta \begin{pmatrix} n \\ i \end{pmatrix} [\int_0^{0.5} g(x,\gamma) f(x,i,n) dx \\
&+ \int_{0.5}^1 g(x,\gamma) f(x,i,n) dx] \\
&= \theta \begin{pmatrix} n \\ i \end{pmatrix} [\int_0^{0.5} g(x,\gamma) f(x,i,n) dx \\
&+ \int_0^{0.5} g(1-x,\gamma) f(1-x,i,n) dx] \\
&= \theta \begin{pmatrix} n \\ i \end{pmatrix} \int_0^{0.5} g(x,\gamma)[f(x,i,n) + f(1-x,i,n)] dx \\
&= 0
\end{aligned}$$

Therefore, $E[Z_i|\gamma] = E[Z_i| - \gamma]$ for $i \neq \frac{n}{2}$.

BIBLIOGRAPHY

[1] Abramowitz, M. and Stegum, IA 1974 Handbook of Mathematical Functions. *New York, Dover.*

[2] Ashburner, M. 1989 Drosophila: A Laboratory Manual. *Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.*

[3] Barton, N.H. 2000 Genetic hitchhiking. Philos. *Trans. R. Soc. Lond. Ser. B* 355:1553- 62.

[4] Bouffard, G. G., J. R. Idol, V. V. Braden, L. M. Iyer, A. F. Cunningham *et al.* 1997 A physical map of human chromosome 7: an integrated YAC contig map with average STS spacing of 79kb. *Genome Res.* 7: 673-692.

[5] Bustamante, C. D., J. Wakeley, S. Sawyer and D. L. Hartl 2001 Directional selection and the site-frequency spectrum. *Genetics* 159: 1779-1788.

[6] Bustamante, C. D., R. Nielsen and D. L. Hartl 2003 Maximum likelihood and Bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data. *Theor. Popul. Biol.* 63 (2): 91-103.

[7] Braverman, M., R. Hudson, N. Kaplan, C. Langley, and W. Stephan. 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* 140: 783796.

[8] Comerson, J. M., and M. Keritman, 2002 Population, evolutionary and genomic consequences of interference selection. *Genetics* 161: 389-410.

[9] Crawford, D.C. et al. 2004 Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat. Genet.* 36: 700-706.

[10] Donnelly and Tavaré S. 1995 Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.,* 29: 401-421.

[11] Ewens, W.J. 1974 A note on the sampling theory for infinite alleles and infinite sites models. *Theor. Popul. Biol.* 6: 143-148.

[12] Fay, J. C., and C. -I Wu 2000 Hitchhiking under positive Darwinian selection. *Genetics* 155: 1405-1413.

[13] Fearnhead, P. and Donnelly, P. 2001 Estimating recombination rates from population genetic data. *Genetics* 159: 1299-1318.

[14] Fearnhead, P., and P. Donnelly 2002 Approximate likelihood methods for estimating local recombination rates. *J. R. Stat. Soc. Ser.* 64: 657-680.

[15] Fearnhead, P., Harding, R. M., Schneider, J.A., Myers, S. and Donnelly, P. 2004 Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots. *Genetics* 167: 2067-2081.

[16] Fearnhead, P. and Smith, N.G.C. 2005 A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes. *Am. J. Hum. Genet.* 77: 781-794.

[17] Felsenstein, J. 1974 The evolutionary advantage of recombination. *Genetics* 78: 737-756.

[18] Fisher, R. A. 1930 The Genetical Theory of Natural Selection. *Oxford University Press, Oxford.*

[19] Fu, Y. -X., and W. -H. Li 1993 Statistical tests of neutrality of mutations. *Genetics* 133: 693-709.

[20] Fu, Y. X. 1994 Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. *Genetics* 138:1375-86.

[21] Fu, Y. X. 1995 Statistical Properties of Segregating Sites. *Theoretical Population Biology* 48: 172-197.

[22] Galtier, N., F. Depaulis and H. N. Barton 2000 Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics* 155: 981-987.

[23] Griffiths, R., and P. Marjoram 1996 Ancestral inference from samples of DNA sequences with recombination. *J. Comp. Biol.* 3: 479-502.

[24] Hartl, D. L., E. N. Moriyama and S. A. Sawyer 1994 Selection intensity for codon bias. *Genetics* 138: 227-234.

[25] Hey, J., and J. Wakeley 1997 A coalescent estimator of the population recombination rate. *Genetics* 145: 833-846.

[26] Hill, W. G., and A. Robertson 1966 The effect of linkage on the limits to artificial selection. *Genet. Res.* 8: 269-294.

[27] Hudson, R. R. 1983 Properties of a neutral allele model with intragenic recombination. *Theor. Pop. Biol.* 23: 183-201.

[28] Hudson, R. R. 1985 The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics* 109: 611-631.

[29] Hudson, R. R., and N. L. Kaplan 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111: 147-164.

[30] Hudson, R. R. 1987 Estimating the recombination parameter of a finite population model without selection. *Genet. Res.* 50: 245-250.

[31] Gene Genealogies and the Coalescent Process. *Oxford University Press, Oxford.*

[32] Hudson, R. R. 2001 Two-locus sampling distributions and their application. *Genetics* 159: 1805-1817.

[33] Hudson, R. R. 2002 Generating samples under a Wright-Fisher neutral model. *Bioinformatics* 18: 337-338.

[34] Innan, H., and Y. Kim 2004 Pattern of polymorphism after strong artificial selection in a domestication event. *Proc. Natl. Acad. Sci. USA* 101 (29): 10667-10672.

[35] Jeffreys, A.J., Murray, J. and Neumann, R. 1998 High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. *Mol. Cell* 2: 267-273.

[36] Jeffreys, A.J., Ritchie, A. and Neumann, R. 2000 High-resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. *Hum. Mol. Genet.* 9: 725-733.

[37] Jeffreys, A.J., Neumann, R., Panayi, M., Myers S. and Donnelly P. 2005 Human recombination hot spots hidden in regions of strong marker association. *Nature Genetics* 37: 601-606.

[38] Jensen, J. D., Y. Kim, V. Bauer-Dumonth, C. F. Aquadro and C. D. Bustanmante 2005 Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* 170: 1401-1410.

[39] Johnson, N., Kotz, S. and Kemp, A.W. 1992 Univariate Discrete Distributions. *John Wiley and sons, NY.*

[40] Kaplan, N. L., Hudson, R.R. and Langley, C.H. 1989 The Hitchhiking Effect Revisited. *Genetics* 123: 887-899.

[41] Kendall, M., 1987 Kendall's Advanced Theory of Statistics. *Oxford University Press, Oxford.*

[42] Kim, Y., and W. Stephan 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160: 765-777.

[43] Kim, Y., and R. Nielsen 2004 Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167: 1513-1524.

[44] Kingman, J.F.C. 1982 The Coalescent. *Stochastic Processes and Their Applications*, 13: 235-248.

[45] Kingman, J. F. C. 1982 Exchangeability and the evolution of large populations. In *Exchangeability in Probability and Statistics,* G. Koch and F. Spizzichino, eds. North-Holland Publishing Company, Amsterdam, pp. 97-112.

[46] Kingman, J. F. C. 1982 On the genealogy of large populations. In *Essays in Statistical Science: Papers in Honor of P. A. P. Moran,* J. Gani and E. J. Hannan, eds. Applied Probability Trust, Sheffield, pp. 27-43. J. Appl. Prob., special volume 19A.

[47] Kocherlakota, S. and Kocherlakota, K. 1992 Bivariate discrete distributions. *New York, Marcel Dekker.*

[48] Kohn, M.H. et al. 2000 Natural selection mapping of the warfarin-resistance gene. *Proc. Natl. Acad. Sci. U.S.A.* 97: 7911-7915.

[49] Kong, A. et al. 2002 A high-resolution recombination map of the human genome. *Nature Genetics* 31(3): 241-7.

[50] Krone, S. M., and C. Neuhauser 1997 Ancestral processes with selection. *Theor. Popul. Biol.* 51: 210-237.

[51] Kuhner, M. K., J. Yamato and J. Felsenstein 1999 *RECOMBINE*, Version 1.0 (http://evolution.genetics.washington.edu/lamarc.html).

[52] Kuhner, M. K., Beerli, P., Yamato, J. and Felsenstein, J. 2000 Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* 156: 439-447.

[53] Lenormand, T. et al. 1998 Evaluating gene flow using selected markers: a case study. *Genetics* 149: 1383-1392.

[54] Lewontin, R. C. 1974 The Genetic Basis of Evolutionary Change. *Columbia Unveristy Press, New York.*

[55] Li, N., and M. Stephens 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165: 2213-2233.

[56] Maynard S. J. and Haigh J. 1974 The hitch-hiking effect of a favorable gene. *Genet. Res.* 23: 23-35.

[57] Morin P.A., Luikart G., Wayne, R.K. and the SNP workshop group 2004 SNPs in ecology, evolution and conservation. *TRENDS in Ecology and Evolution* 19(4):208-215.

[58] McVean, G., P. Awadalla and P. Fearnhead 2002 A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160: 1231-1241.

[59] McVean, G., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley *et al.* 2004 The fine-scale structure of recombination rate variation in the human genome. *Science* 304 (5670): 581-584.

[60] Nagaraja, R., S. Macmillan, J. Kere, C. Jones, S. Griffin *et al.* 1997 X chromosome map at 75-kb STS resolution, revealing extremes of recombination and GC content. *Genome Res.* 7: 210-222.

[61] Nelder, J. A. and Mead, R. 1965 "A Simplex Method for Function Minimization." *Comput. J.* 7, 308-313.

[62] Neuhauser, C., and S. M. Krone 1997 The genealogy of samples in models with selection. *Genetics* 145: 519-534.

[63] Nielsen, R. 1998 Maximum Likelihood Estimation of Population Divergence Times and Population Phylogenies under the Infinite Sites Model. *Theor. Pop. Biol.* 53: 143-151.

[64] Nielsen, R. 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* 154: 931-942.

[65] Nielsen, R. 2001 Statistical tests of selective neutrality in the age of genomics. *Heredity* 86: 641-647.

[66] Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, and C. D. Bustamante. 2005 Genomic scans for selective sweeps using SNP data. Genome Research. *Genome Research* 15:1566-1575.

[67] Nordborg, M. 2001 "Coalescent theory" Chapter 7 in D. Balding, M. Bishop, and C. Cannings, editors, *Handbook of Statistical Genetics*. Wiley, Chichester, UK.

[68] Rannala, B. and M. Slatkin. 2000 Methods for multipoint disease mapping using linkage disequilibrium. *Genetic Epidemiology* 19: S71-S77

[69] Robertson, A. 1961 Inbreeding in artificial selection programmes. *Genet. Res.* 2: 189-194.

[70] Sawyer, S A, D E Dykhuizen, R DuBose, L Green, T Mutangadura-Mhlanga, D Wolczyk, and D. L. Hartl 1987 Distribution and abundance of insertion sequences among natural isolates of Escherichia coli. *Genetics* 115: 51-63.

[71] Sawyer, S. A., and D. L. Hartl 1992 Population genetics of polymorphism and divergence. *Genetics* 132: 1161-1176.

[72] Sawyer, S. A., R. J. Kulathinal, C. D. Bustamante and D. L. Hartl 2003 Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. *J. Mol. Evol.* 57: S154-S164.

[73] Schlotterer, C. 2002 Towards a molecular characterization of adaptation in local populations. *Curr. Opin. Genet. Dev.* 12: 683 687.

[74] Simonsen K. L., G. A. churchill, and C. F. Aquadro 1995 Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* 141: 413-429.

[75] Slade, P. F. 2001 Simulation of 'hitch-hiking' genealogies. *J. Math. Biol.* 42 (1): 41-70.

[76] Slatkin, M., and R. R. Hudson 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129: 555-562.

[77] Stephen, M. R. and F. J. Ayala, 1998 The Recent Origin of Allelic Variation in Antigenic Determinants of Plasmodium falciparum. *Genetics* 150: 515-517.

[78] Tajima, F. 1983 Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105: 437-460.

[79] Tajima, F. 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585-595.

[80] True, J. R., J. M. Mercer and C. C. Laurie 1996 Differences in crossover frequency and distribution among three sibling species of Drosophila. *Genetics* 142: 507-523.

[81] Wakeley, J. 1997 Using the variance of pairwise differences to estimate the recombination rate. *Genet. Res.* 69: 45-48.

[82] Wakeley, J. 2001 The coalescent in an island model with variation among demes. *Theor. Popul. Biol.* 59: 133-144.

[83] Wakeley, J. 2003 Polymorphism and divergence for island-model species. *Genetics* 163: 411-420.

[84] Wakeley, J. 2004 Metapopulation models for historical inference. *Mol. Ecol.* 13:865-875.

[85] Wall, J. D. 2000 A comparison of estimators of the population recombination rate. *Mol. Biol. Evol.* 17: 156-163.

[86] Wall, J. D. 2004 Estimating recombination rates using three-site likelihoods. *Genetics* 167: 1461-1473.

[87] Watterson, G. A. 1975 On the number of segregating sites in genetic models without recombination. *Theor. Popul. Biol.* 7: 256-276.

[88] Wayne, M. L and K. Simonsen 1998. Statistical tests of neutrality in the age of weak selection. *Trends in Ecology and Evolution* 13: 236-240.

[89] Wright, S. 1931 Evolution in mendelian populations. *Genetics*, 16:97-159.

[90] Wright, S. 1938 The distribution of gene frequencies under irreversible mutation. *Proc. Natl. Acad. Sci. USA* 24: 253-259

[91] Williamson, S 2003 Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. *Biol. Evol. Mol. Biol. Evo.* 20: 1318-1325.

[92] Williamson, S., A. F. Alon and C. D. Bustamante 2004 Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. *Genetics* 168: 463-475.

[93] Williamson, S., R. Hernandez, A. F. Alon, L. Zhu, R. Nielsen *et al.* 2005 Simultaneous inference for selection and population growth from patterns of variation in the human genome. *Proc. Natl. Acad. Sci. USA* 102 (22): 7882-7887.

[94] Yang, Z. 1997 On the estimation of ancestral population sizes of modern humans. *Genetical Research Cambridge* 69:111-116.

[95] Zhu, L. and C. D. Bustamante 2005 A composite likelihood approach for detecting directional selection from DNA sequence data. *Genetics* 170: 1411-1421.